UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

E.A.P DE INGENIERÍA DE SISTEMAS

Desarrollo de un data warehouse para la difusión de los censos nacionales del Instituto Nacional de Estadística e Informática de Perú

Tesina

Para optar el Título de Ingeniero de Sistemas

AUTORES

Norma Arcadia Gutiérrez Peña

Juber Raúl Palomino Campos

LIMA – PERÚ

2013

FICHA CATALOGRAFICA

GUTIERREZ PEÑA, Norma Arcadia
PALOMINO CAMPOS, Juber Raúl

DESARROLLO DE UN DATA WAREHOUSE PARA LA DIFUSION DE LOS CENSOS NACIONALES DEL INSTITUTO NACIONAL DE ESTADISTICA A INFORMATICA DE PERU

BASE DE DATOS
Data Warehouse e Inteligencia de Negocios
Lima, Perú 2013

Tesina, Facultad de Ingeniería de Sistemas e Informática, Pregrado, Universidad Nacional Mayor De San Marcos

Formato 28 x 20 cm. Páginas 101

DEDICATORIA:

Dedicamos el presente trabajo de tesina a nuestras familias, por confiar siempre en nosotros y brindarnos su apoyo para lograr nuestros objetivos a nivel personal y profesional.

AGRADECIMIENTOS

Agradecemos a los profesores de la facultad de Ingeniería de Sistemas e Informática por la formación brindada, al profesor Luis Alarcón Loayza por la asesoría en el desarrollo de este trabajo y a los docentes del Comité Evaluador por sus aportes. También nuestro reconocimiento al Instituto Nacional de Estadística e Informática del Perú, por brindarnos las facilidades para la aplicación de nuestros conocimientos en la conducción de proyectos estratégicos para dicha institución así como por el apoyo al desarrollo del proyecto materia de la presente tesina.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA
ESCUELA ACADEMICO PROFESIONAL DE INGENIERIA DE SISTEMAS

DESARROLLO DE UN DATA WAREHOUSE PARA LA DIFUSION DE LOS CENSOS NACIONALES DEL INSTITUTO NACIONAL DE ESTADISTICA A INFORMATICA DE PERU

Autores: GUTIERREZ PEÑA, Norma Arcadia
PALOMINO CAMPOS, Juber Raul

Asesor: ALARCON LOAYZA, Luis

Título: Tesina, para optar el Título Profesional de Ingeniero de Sistemas
Fecha: Febrero del 2013

RESUMEN

El proyecto tiene por objetivo el desarrollo de un Data Warehouse de los Censos Nacionales para el INEI de Perú, que facilite el acceso a la información censal a los diversos usuarios que requieran conocer la situación socio demográfico del país y puedan tomar decisiones. Para ello se aplicarán los enfoques y técnicas para manejo de información estadística y utilización de herramientas tecnológicas.

La propuesta surge a raíz de que el INEI cuenta con diversos aplicativos informáticos a través de los cuales muestra la información censal en forma de cuadros estadísticos, bases de datos planas y resúmenes a diversos niveles de agrupación, las que no están integradas y no cuentan con información descriptivas única sobre sus características, todo ello dificulta que los usuarios puedan encontrar información. En consecuencia hay una demanda insatisfecha sobre la información censal que provee el INEI.

Palabra claves: Censo, Data Warehouse, Data Mart, Inteligencia de Negocios

MAJOR NATIONAL UNIVERSITY OF SAN MARCOS

FACULTY OF SYSTEMS ENGINEERING AND INFORMATICS
ACADEMIC PROFESSIONAL SCHOOL OF SYSTEMS ENGINEERING

DESARROLLO DE UN DATA WAREHOUSE PARA LA DIFUSION DE LOS CENSOS NACIONALES DEL INSTITUTO NACIONAL DE ESTADISTICA A INFORMATICA DE PERU

Author: GUTIERREZ PEÑA, Norma Arcadia
PALOMINO CAMPOS, Juber Raul

Assessor: ALARCON LOAYZA, Luis

Title: Thesis, to obtain the title of Engineer and Computer Systems

Date: February 2013

ABSTRACT

The project aims at developing a data warehouse of the National Census for INEI of Peru, which facilitates access to the census information to the various users who need to know the country's socio demographic and can make decisions. This will apply the approaches and techniques for management of statistical information and use of technology tools.

The proposal stems from the INEI has several computer applications through which census data shows as statistical tables, databases, flat summaries to various levels of aggregation, which are not integrated and do not have only descriptive information about its features, all this makes it difficult for users to find information. Consequently there is an unmet demand on the census information provided by the INEI.

Key Words: Census, Data Warehouse, Data Mart, Business Intelligence

ÍNDICE

INDICE DE FIGURAS

INDICE DE TABLAS

INTRODUCCIÓN

Tener un diagnóstico de la situación socio demográfico y económica de un país es la base para la evaluación, definición, planificación y ejecución de programas y políticas tanto públicas y privadas, a fin de mejorar las condiciones de vida de sus habitantes y por ende el desarrollo de un País. Para ello se requiere de información detallada, actualizada y con calidad sobre los aspectos estructurales del país.

En este contexto, los Censos Nacionales de Población y Vivienda realizado por el Instituto Nacional de Estadística e Informática (INEI) tienen un papel primordial, al constituirse en la única fuente de datos capaz de brindar información de todas y cada una de las personas y las viviendas del país. Pero pierde su valor e importancia si no está integrada o no está a disponibilidad para todo usuario que lo requiera. Por ello, el INEI con apoyo de organismos internacionales viene orientando sus esfuerzos a mejorar la difusión de la información estadística que produce, dentro del contexto de modernización del estado peruano.

El propósito de este estudio es Desarrollar un Data Warehouse de los Censos Nacionales para el INEI de Perú, que facilite el acceso a la información, para así responder a todas las necesidades de información censal y que sirva de base para la toma de decisiones estratégicas y tácticas de usuarios en las entidades públicas o privadas. Su desarrollo ha contado con el apoyo de la Jefatura del INEI y el diseño ha sido tomando como referencia soluciones similares en los institutos de estadística de México y España.

Este trabajo consta de cinco capítulos: en el primer capítulo se presenta el planteamiento del problema, la justificación, su objetivo entre otros puntos. En el segundo capítulo se

presenta el marco referencial, donde se trata sobre la relevancia de información para toma de decisiones, data warehouse, inteligencia de negocios y censos nacionales. En el tercer capítulo se presenta el estado de arte, donde se muestran aplicaciones similares, metodología de desarrollo e implementación de un data warehouse. En el cuarto capítulo se presenta el desarrollo del data warehouse siguiendo la metodología de Barquín Internacional. Finalmente en el quinto capítulo se presentan las conclusiones y recomendaciones.

CAPÍTULO I: PLANTEAMIENTO METODOLÓGICO

1.1 Antecedentes

1.1.1 Necesidad de Información Estadística

Actualmente las organizaciones tienen claro la importancia de la gestión de la información y las ventajas competitivas que implica su uso. Este proceso de gestión consiste en lograr de una manera eficiente el análisis de los datos que se producen en las organizaciones y su entorno, transformándola en un recurso estratégico de gran importancia que facilita tomar mejores decisiones, lograr ventaja competitiva, asegurar la calidad y planificar el futuro haciendo uso de las Tecnologías de la Información y Comunicaciones (TIC).

Por ello, es necesario que las organizaciones cuenten con un entorno de análisis de datos que este orientado a la producción de conocimiento, y poner a disposición de los usuarios, herramientas que pueden utilizar para mejorar su productividad en el desempeño de sus funciones. Dicho entorno debe ser independiente del entorno operativo y debe contar con una estructura apropiada de almacenamiento de datos, que permita obtener en forma rápida respuestas a las consultas recorriendo miles o millones de registros.

En la implementación de este entorno destacamos tres elementos principales: (i) un Data Warehouse (DW) como la gran base de datos corporativa, (ii) los procesos de

transformación y alimentación de esta base de datos y (iii) conjunto de herramientas que permite al usuario final acceder a la información corporativa de acuerdo a sus necesidades.

Por otro lado, es notorio que en los últimos años en Perú, específicamente en el sector gobierno, se ha venido produciendo un impresionante desarrollo de las tecnologías de la información y las comunicaciones (TIC). Su introducción y aplicación masiva a gran velocidad han sentando las bases de una nueva cultura, la de la sociedad de la información. La creciente necesidad de información para la aplicación de políticas públicas, ha potenciado este desarrollo y a la vez ha determinado un cambio radical en las instituciones públicas.

Dentro de este contexto, la información estadística oficial producida por el Instituto Nacional de Estadística e Informática (INEI) de Perú, da respuesta a muchas de las necesidades de información que la sociedad requiere e influye notablemente en la toma de decisiones en el sector privado y público. Para ello, el INEI cuenta con los censos nacionales como principal fuente de información estadística, y es una poderosa herramienta para la recolección de datos que permite conocer la realidad socio demográfico, económico y agrario del país.

1.1.2 Presentación de la Institución

La coyuntura macroeconómica que hoy vive el Perú, constituye una gran ventana de oportunidades para poner en marcha el desarrollo que el país espera, a través de políticas adecuadas que aseguren un crecimiento sostenido y equitativo para que los beneficios lleguen a toda la población. El gran desafío que se ha impuesto nuestro país es acortar la brecha entre quienes más y menos tienen. Para ello, demanda información estadística como un soporte eficaz para apoyar el diseño y formulación de proyectos públicos, cuya calidad y pertinencia puedan cambiar sensiblemente el desempeño del país.

Los censos, permiten conocer con mayor exactitud cuántos y cómo somos, cuales son las principales necesidades en servicios públicos, infraestructura, salud, educación, identificar con precisión los grupos vulnerables y sus características; conocer las

oportunidades de crecimiento productivo; de este modo, la estadística censal se convierte en uno de los pilares más importantes para alcanzar un desarrollo equitativo, sostenible y democrático del país.

En este contexto, el **INEI**, en su condición de órgano rector del Sistema Estadístico Nacional (SEN) de Perú, busca la generación de información para la implementación de programas y proyectos públicos y no públicos, que impacten en el campo económico y social del territorio nacional.

La **visión y misión del INEI**, esta expresada en los términos siguientes: [1]

Visión:

"*En el año 2012, el Perú cuenta con un ágil y eficiente sistema nacional de coordinación, producción y difusión de información estadística confiable, oportuna y de calidad, con cobertura de datos desagregada a todo nivel político-administrativo, que contribuye eficazmente al diseño, implementación y evaluación de políticas públicas, programas y proyectos de desarrollo que impactan en el crecimiento económico, reducción de la pobreza y conservación ambiental. Satisface plenamente los requerimientos de los usuarios del sector público y privado. La información estadística es de fácil acceso y su producción y difusión se realiza con el uso intensivo de la tecnología de información más avanzada*."

Misión:

"*El Sistema Estadístico Nacional es la red de entidades del sector público a nivel central, regional y local que, bajo la rectoría del Instituto Nacional de Estadística e Informática - INEI, produce y difunde información estadística oficial, en forma integrada, coordinada, racionalizada y bajo una normatividad técnica común, con el propósito de contribuir al diseño, monitoreo y evaluación de políticas públicas y al proceso de toma de decisiones de los agentes socio económicos y de la comunidad académica, con estadísticas oportunas, confiables y de calidad*."

[1] Pagina Web del INEI http://www.inei.gob.pe/

El objetivo estratégico que guía el comportamiento operacional y táctico de la institución es: "*Producir y difundir información estadística confiable, oportuna y de calidad, de fácil acceso, que satisfaga los requerimientos de los usuarios*."[2]

Para el cumplimiento de sus funciones, el INEI cuenta con la Estructura Orgánica que se muestra en la Figura 1.

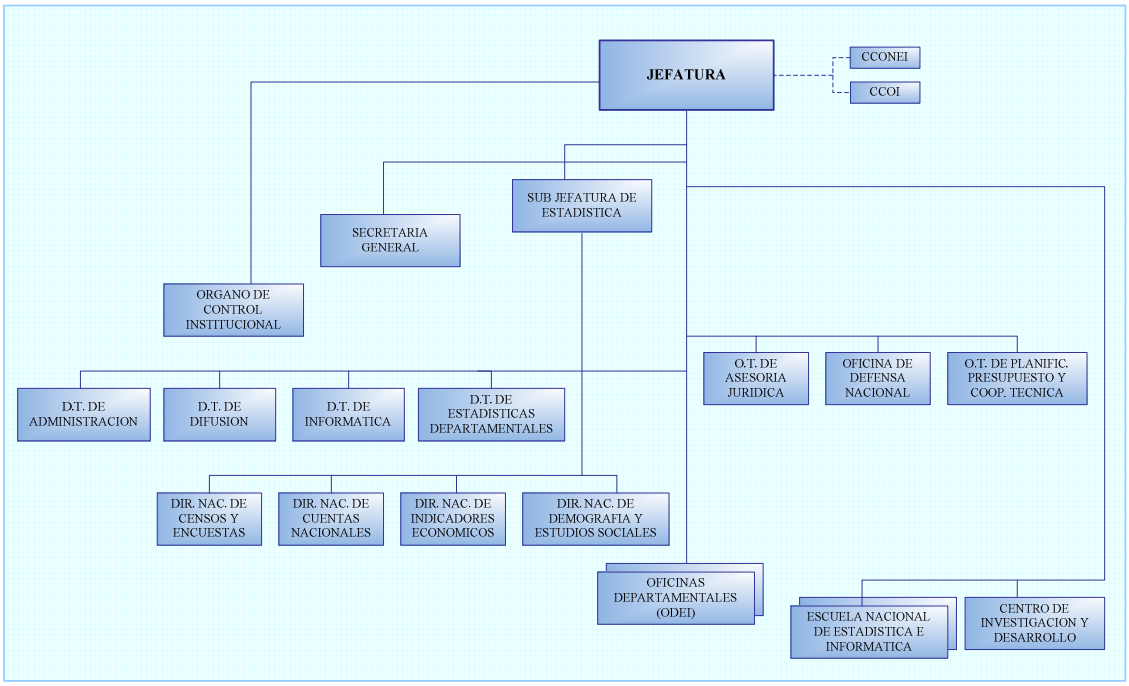

Figura 1: Estructura Orgánica del INEI

Fuente: INEI (2012)

1.2 Definición del Problema

La amplia gama de información censal producida por el INEI a través de los años, no está integrada en una sola base de datos (están en diferentes formatos de archivos: Oracle, SQL Server, Postgress, Texto). Esto representa un problema para los analistas de información del INEI, quienes tienen dificultades para acceder a la información estadística que producen.

Las bases de datos son consultadas a través de los diversos aplicativos que se desarrollan para la difusión de la información censal. Estos presentan la información en

[2] Plan Estratégico Nacional para el Desarrollo Estadístico 2008 -2012, DS Nº089-2007-PCM

formatos preestablecidos como cuadros y gráficos estadísticos o cuadros en Excel; generando algunas veces resultados diferentes para una misma variable por no tener definido su metadato (descripción técnica de la variable).

Los usuarios externos, tienen requerimientos de información estadística que son solicitados por correo electrónico o documento físico (oficio, carta, memorando), lo cual genera una serie de trámites administrativos y una sobrecarga de trabajo para su atención en las oficinas de Informática y de Difusión del INEI, quienes muchas veces tienen que realizar procesos repetitivos para generar la información solicitada y en algunas oportunidades con errores en los resultados. Existen casos en las cuales simplemente no se les atiende porque la solicitud no especifica claramente los datos que está solicitando o en su defecto es demasiado laborioso procesarla. Esto genera en el usuario malestar por la espera de la atención a su solicitud.

En consecuencia, el enunciado del problemas es: *La existencia de diversos aplicativos informáticos a través de los cuales el INEI difunde la información de los censos nacionales en forma de cuadros estadísticos, bases de datos planas y resúmenes a diversos niveles de agrupación no están totalmente articuladas; generando sobrecarga laboral, demora en atención a las solicitudes de información, posibles diferencias en la información y malestar en los usuarios*.

1.3 Objetivos

1.3.1 Objetivo General

Desarrollar un Data Warehouse de los censos nacionales para el INEI de Perú, que facilite el acceso a la información censal a los diversos usuarios en el menor tiempo posible.

1.3.2 Objetivos Específicos

- Identificar las necesidades de información de los diversos tipos de usuarios para la mejor organización de la información.

- Comparar las metodologías más utilizadas para la construcción del DW e identificar la más adecuada para el manejo de información censal.

- Diseño e implementación de un DW que permita la integración de los diversos censos nacionales.

- Diseño de procesos de extracción, transformación y carga de información a la base de datos multidimensional.

- Comprobar que el DW permitirá reducir el tiempo de acceso a la información de los censos nacionales.

1.4 Justificación

El proyecto surge como apoyo al proceso de reforma de los programas de superación de la pobreza y desarrollo de capital humano.

Las razones que justifican el desarrollo de un DW de los censos nacionales, en lugar de tener aplicativos ad-hoc accesando directamente a las bases de datos operacionales son:

- **Integración de la información de los censos realizados por el INEI**
  La existencia de bases de datos de múltiples censos nacionales (población, Vivienda, Económico, Agropecuario, Universitario, Comunidades Indígenas, Pesca, etc.) cuyas estructuras y convenciones de nombres no están totalmente estandarizadas o no están totalmente documentadas, dificulta la tarea de las personas encargadas de tomar decisiones en base a dicha información. Por tanto es fundamental que se desarrolle un almacén integrado de datos definitivos con información estadística obtenida de los diversos censos nacionales, que ayude a articular la información de todo el sistema estadístico peruano para tener una visión completa del país.

- **Facilitar a los usuarios el acceso a la información censal**
  Durante los últimos años el INEI ha realizado esfuerzos en la construcción y difusión de diversos aplicativos informáticos ad-hoc mostrando información censal a nivel macro (nacional, departamental, provincial y distrital). Estos

aplicativos no han cubierto totalmente las necesidades de información de los usuarios porque se limitan a consultar información estática en cuadros estadísticos pre elaborados o en publicaciones físicas que dificulta la generación de nuevos indicadores. Los usuarios requieren mayor nivel de detalle de la información (centros poblados, zonas y manzanas) y sobre todo la posibilidad de generar ellos mismos sus indicadores en el menor tiempo posible. Lo cual es atendido con la facilidad de navegación o drilling de los datos que proporciona un DW.

- **Mayor nivel de limpieza de los datos censales**

  Por el gran volumen de la información censal, los procesos de consistencia y corrección se realizan a través de procesos batch a nivel distrital; mediante el cual se asignan valores obtenidos en función a diversas técnicas estadísticas a los campos detectados con "error" o "vacías". Algunas inconsistencias no se detectan a nivel macro (nacional, departamental, provincial, distrital) pero cuando se baja a nivel de detalle (centros poblados, zonas o manzanas) podrían existir inconsistencias; lo cual se corrige en la limpieza realizada por los programas de ETL (extracción, transformación y carga) que se encarga de asegurar que los datos sean consistentes antes de ser cargados dentro del DW.

- **Disponibilidad de Datos Históricos de los Censos**

  Los datos históricos de los diversos censos nacionales por lo general no están disponibles en los sistemas operacionales, pero son fundamentales para un análisis que permita conocer las tendencias y cambios en los diversos sectores del País a través del tiempo.

- **Disponibilidad de datos agregados**

  Por lo general, para la toma de decisiones con información censal, no hay necesidad de ingresar en tiempo real a nivel detallado o microdato, por lo cual se requiere la disponibilidad de información agregada con la posibilidad de un análisis multidimensional.

- **Reducir tiempo de acceso a la información censal**

Si el usuario requiere algún cuadro estadístico a la medida de sus necesidades y no está disponible en las publicaciones o aplicativos difundidos por el INEI; entonces debe realizar su solicitud de información siguiendo protocolos establecidos para tal fin, lo cual podría tomar 12 días o más como se muestra en Figura 2. Al implementar un DW en internet se abre la posibilidad de que el usuario construya sus propias consultas sin tener que recurrir a terceros.

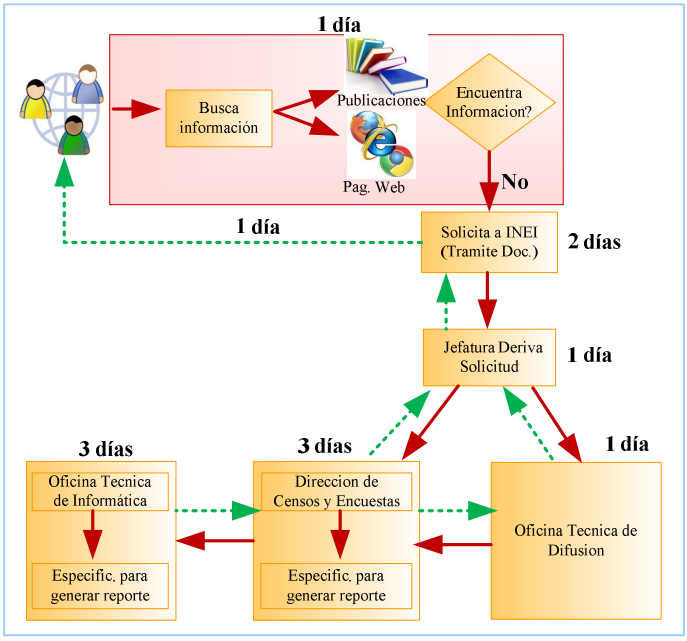

Figura 2: Tiempo estimado para obtener información

Fuente: Elaboración Propia

- **Seguridad de la Información Censal**

  Los datos de los censos nacionales es un activo del país, por lo que es necesario contar con mecanismos que permitan protegerlos de la modificación, destrucción o divulgación no autorizada; así como la ocurrencia de estos eventos de modo accidental. Lo cual se garantiza ya que la información de los DW no son actualizados.

- **Mejor aprovechamiento del recurso invertido por el estado en proyectos censales**

  Todos los países requieren conocer el número exacto de habitantes, sus características, la calidad de vida de la población, su ubicación y sus necesidades. Así mismo, requieren conocer diversos indicadores económicos y

sociales, con lo cual se puede diseñar políticas, programas y proyectos para mejorar las condiciones de vida de los habitantes; así como proyectar el crecimiento poblacional de acuerdo a la división política administrativa del país.

Para obtener dicha información, se realizan los censos nacionales de población y vivienda; en el caso de Perú, estos son realizados cada 10 años en cumplimiento de la ley General de los Censos Nacionales y su ejecución está a cargo del INEI. Para el último censo de población y vivienda ejecutado el 21 de octubre del 2007 se invirtió 63`035,739 de nuevos soles [3], se aplicó un cuestionario de 63 preguntas y se captó información de 27`419, 294 habitantes[4], de donde se deduce que el costo por habitante es de aproximadamente 2.3 nuevos soles y el costo estimado de cada pregunta es aproximadamente de 1'000, 567 de nuevos soles.

Como se observa, la inversión es bastante elevada y es evidente que este tipo de información estadística es vital para el país, por tanto es parte de la infraestructura básica de toda sociedad, al constituirse en una valiosa herramienta para lograr una sociedad informada y participativa por lo cual es fundamental poner en internet toda la información censal con facilidades para el realizar el análisis multidimensional.

Por las razones planteadas y para atender estas necesidades de información censal, se desarrolla el DW de los Censos Nacionales y así facilitar a los usuarios el acceso a los grandes volúmenes de información. De este modo se incorporara a los procesos tradicionales del INEI un nuevo enfoque de almacenamiento, organización, difusión y gestión de la información considerando el tiempo, la geografía y diferentes temáticas a todo nivel de detalle.

También constituye un gran avance para el Sistema Estadístico Nacional (conformada por todas las entidades gubernamentales de Perú) para que la experiencia, los estándares y normas que se logren con el desarrollo del DW, permitan contar con información estadística de calidad y de fácil acceso para la sociedad.

---

[3] Presupuesto del Censo Nacional mostrado en Pagina web del INEI

[4] Primeros Resultados de los Censos Nacionales 2007

Es importante resaltar que esta solución posiciona al INEI como una entidad clave para brindar información estratégica porque a nivel latinoamericano solo México y Brasil cuentan con una solución similar.

1.5 Alcance y Limitaciones

1.5.1 Alcances del Estudio

De acuerdo a los objetivos generales y específicos planteados, el presente proyecto tiene como tarea el desarrollo de un DW para integrar la información de los Censos Nacionales y facilitar el acceso de los usuarios al mismo.

Las fuentes de información a utilizarse inicialmente son las bases de datos correspondientes a los datos captados con el formulario del último Censo Nacional de Población y Vivienda realizada en el año 2007, la cual ha sido validada a todos los niveles geográficos (Nacional, Departamental, Provincial, Distrital, Centro Poblado, Zonas y Manzanas) y según cada alternativa de respuesta a las preguntas del formulario censal.

El usuario accederá a la información vía Internet, mediante herramientas que le permitan diversas opciones de visualización (texto, gráficos o exportación de archivo), creación de sus propios indicadores dinámicamente consultando el diccionario de metadatos.

De este modo, busca atender a una diversidad de público que no necesariamente cuenta con experiencia en el uso y análisis de información estadística. Los tipos de usuarios identificados son:

- Tomadores de decisiones del gobierno central, los gobiernos regionales y locales
- Analistas de información de empresas privadas
- Académicos
- Profesionales
- Estudiantes

- Público en general

Se espera brindar información a cerca de 10 millones de usuarios con acceso a Internet con los que cuenta el Perú según Agencia Central de Inteligencia (CIA)[5] y según la Oficina Nacional de Gobierno Electrónico (ONGEI)[6].

1.5.2 Delimitaciones del Problema

El DW de los Censos Nacionales, integra y administra inicialmente la información de los Censos de Población y Vivienda 2007 y el Censo de Comunidades Indígenas. La información a la que accederá el usuario corresponde a los datos básicos y a los niveles geográficos determinados por la política institucional de difusión de datos y resguardo de la confidencialidad de los mismos, los cuales estarán plasmados en el sistema.

El usuario final accede a la información por Internet, por lo tanto está limitado a los lugares que cuentan con este servicio.

El DW se pondrá en producción en la infraestructura Tecnológica que tiene asignado, la institución para el proyecto.

1.6 Propuesta

Para la solución del problema planteado, se propone el diseño e implementación de un DW, que permita lograr la transformación, estandarización e integración de los datos censales en una gran base de datos multidimensional, brindando una herramienta sencilla y rápida para el análisis el cual sirva de base para la toma de decisiones con la oportunidad necesaria.

Los principales elementos utilizados son:

---

[5] Field Listing: Internet de la CIA

[6] ONGEI: Celebro el Día de Internet, 2012

- Procesos automáticos de extracción, transformación y carga para mantener la información actualizada.
- Datos y metadatos, que están validadas, integradas y documentadas.
- Herramientas que facilitan el acceso a la base de datos para la explotación y análisis de información.

1.7 Organización de la Tesina

En el capítulo uno se presenta el planteamiento metodológico, donde se justifica el porqué de la tesina, su objetivo, alcances y limitaciones, entre otros puntos.

En el capítulo dos se presenta el marco teórico referencial, donde se trata sobre la importancia de los datos, DW, inteligencia de negocios y censos nacionales.

En el capítulo tres se presenta el estado de arte, donde se describen soluciones similares, diversas metodologías para la construcción de DW, la comparación entre ellos y la selección de la metodología a aplicar.

En el capítulo cuatro se presenta el desarrollo del DW, donde primero se detalla el planeamiento, análisis e identificación de los requerimientos, diseño del modelo dimensional, luego se explica la presentación de la información mediante la obtención de cuadros estadísticos dinámicos.

En el capítulo cinco se realiza el análisis de los resultados obtenidos, y como se optimizó la tarea de los usuarios.

Finalmente, en el capítulo seis se presentan las conclusiones y recomendaciones.

Capítulo II: MARCO REFERENCIAL

2.1 Datos operacionales y datos informativos

Los procesos automatizados de un negocio utilizan **datos operacionales**, que son los registros resultantes de las transacciones que realiza. Los datos creados durante la ejecución de los procesos, muchas veces contienen valores incorrectos, son muy detallados y no tienen mucho uso en los negocios debido a su gran volumen, ubicación y diversos formatos.

Los **datos operacionales** son específicos para cada aplicación y usualmente son almacenados de manera separada. Estos datos son útiles en la medida en que se aprovechen para satisfacer el proceso de las aplicaciones predefinidas. Mayormente se requieren sólo datos actuales y estos deben ser mantenidos al día haciendo actualizaciones frecuentes en la base de datos.

El usuario necesita para sus actividades de análisis los **datos informativos**, que en realidad son los datos operacionales que han sido modificados, depurados, transformados, consolidados y organizados por temas de negocio, que contengan valores históricos, que se encuentren disponible para análisis durante períodos largos y que sea accesible de manera fácil y flexible.

Los datos operacionales son manejados, por los sistemas operacionales o transaccionales (On Line Transactional Processing, OLTP); mientras que los datos informativos son los que conforman un DW, el cual tiene como fin comprender, medir y

administrar parámetros estratégicos del negocio. En Tabla 1, se muestran las diferencias entre los datos operacionales y los datos informativos.

Datos Operacionales	**Datos Informativos**
Orientados a una aplicación	Orientados a un tema
Integración limitada	Integrados
Constantemente actualizados	Estáticos
Sólo valores actuales	Valores a lo largo del tiempo
Soportan operaciones diarias	Soportan decisiones de administración

Tabla 1: Datos operacionales versus datos informativos

Fuente: Herrera, Cristhian (2007)

2.2 Data Warehouse

Desde mediados de los ochenta, en el entorno empresarial, ha cobrado importancia el concepto Data Warehouse o almacén de datos, como la plataforma que concentra en forma integrada toda la información clave para la organización, sus fuentes de información son las bases de datos corporativas así como fuentes externas.

Una clave del éxito de las organizaciones modernas es el acceso a la información correcta, en el tiempo adecuado, en el lugar correcto y en la forma adecuada, todo esto se hace viable de manera rápida cuando se cuenta con un DW.

2.2.1 Definición

Existen muchas y variadas definiciones sobre DW cada una de las cuales esquematizan diversos aspectos de su implementación. A continuación se presentan, las dos definiciones más representativas[7]:

- **Bill Inmon**

[7] Inteligencia de Negocios, CIBETEC 2011

"El DW es una colección de datos, orientados a un tema, integrados, no volátiles, variantes en el tiempo, organizados para el apoyo a toma de decisiones."

- **Ralph Kimball:**

  "Un DW es una copia de los datos transaccionales, específicamente diseñada para realizar consultas y análisis."

2.2.2 Características

Las características más relevantes del DW según Bill Inmon es que está orientado al tema, es integrado, es de tiempo variante y no es volátil.

- **Orientado al tema**

  Como se muestra en la Figura 3, clasifica la información en base a los aspectos que son de interés para la organización y según la perspectiva del usuario final; no como los sistemas operacionales que organizan sus datos desde la perspectiva de la aplicación que están relacionadas con el diseño de la base de datos y del proceso. En el DW, el modelamiento y el diseño de la base de datos, es considerando la organización de la información por temas de modo que un analista pueda formular las preguntas necesarias excluyendo la información que no será usada.

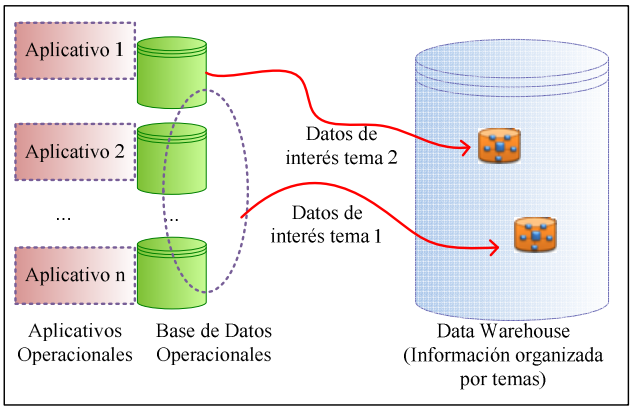

Figura 3: Organización temática del Data Warehouse

Fuente: Basado en CIBERTEC (2011)

- **Integrado**

  Es la característica más importante, pues vincula datos recolectados de diferentes sistemas operacionales de la organización y/o fuentes externas.

  Como se representa en la Figura 4, la información encontrada en el interior de un DW está siempre integrada, pues es un conjunto de datos y metadatos perfectamente vinculadas con respecto al nombre de las variables, formatos de los distintos campos, medida de los atributos, codificación, entre otros.

  Lo que no sucede siempre en los sistemas transaccionales, que han sido desarrollados a través de los años y donde los desarrolladores han tomado sus propias decisiones sobre construir una aplicación, estilos y diseños personalizados que se muestran de muchas maneras.

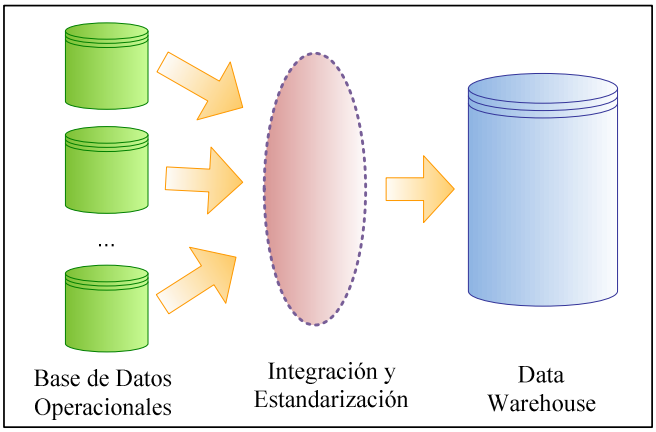

Figura 4: Integración del Data Warehouse

Fuente: Basado en CIBERTEC (2011)

- **De tiempo variante**

  Los datos son relativos a un período de tiempo y deben ser actualizados periódicamente; cada periodo es almacenado como la foto que corresponde a un período de tiempo.

  El tiempo debe estar en todos y cada uno de los registros del DW, pues permite saber en qué momento tenía un valor específico. De este modo, se cuenta con información histórica que permite comparar datos en distintos períodos e identificar tendencias.

Esta característica, es muy diferente para la información encontrada en el ambiente operacional, donde la información que se requiere al momento de acceder, son valores actuales a partir del momento de acceso.

- **No volátil**

  Los datos que son almacenados en el DW no sufren actualización, sólo permite adicionar nuevos datos y acceder a los ya almacenados que son guardados como datos históricos.

  La información es útil sólo cuando es estable, característica esencial para el análisis y la toma de decisiones, lo cual no sucede con los datos operacionales pues cambian de momento a momento.

  En el diseño físico del DW, se pueden tomar libertades para optimizar el tiempo de acceso a los datos, particularmente al usar la normalización o desnormalización física.

2.2.3 Estructura Lógica

Los DW tienen una estructura lógica particular y la información se organiza con diversos niveles de detalle o granularidad de información. Estos niveles o tipos de datos son: Detalle de datos actuales, Detalle de datos antiguos, Datos ligeramente resumidos, Datos completamente resumidos y el Metadato.[8]

2.2.4 Arquitectura

Permite comprender cómo se relacionan todos los componentes involucrados en una solución de DW, indicando el orden en el que se deben hacerse las cosas, como se muestra en la Figura 5.

[8] Hefesto-v2 Metodología para la construcción de un DWH

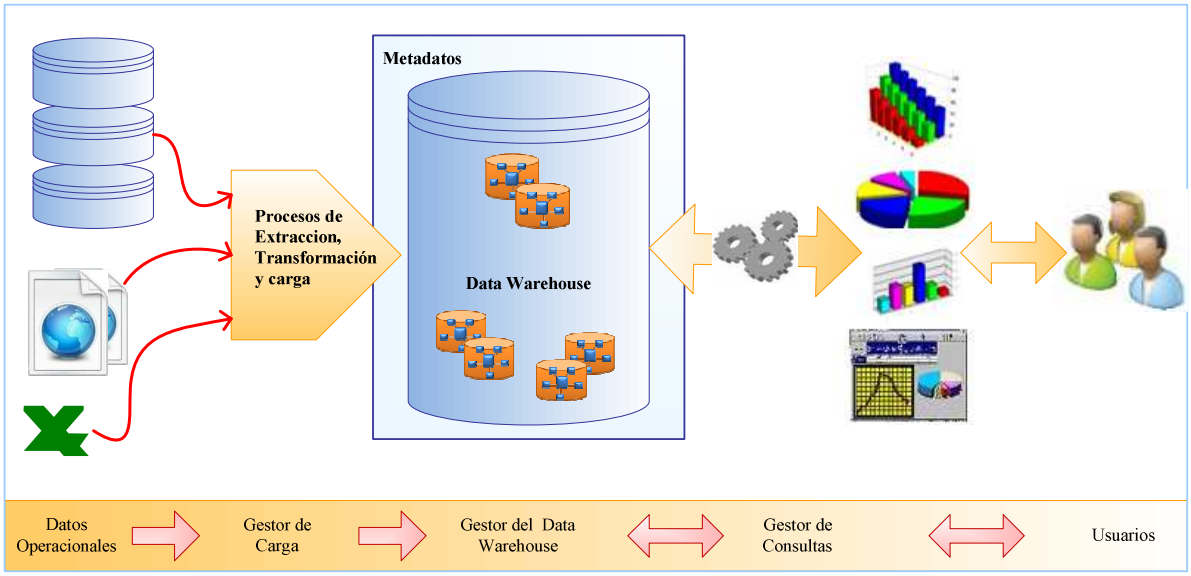

Figura 5: Arquitectura del Data Warehouse

Fuente: Basado en Bernabeu, Ricardo (2010)

Donde:

- **Datos operacionales**

  Representa toda aquella información transaccional que genera la organización al realizar sus procesos operativos, así como datos de fuentes externas que son de su interés y puede disponer. Los formatos en que se encuentran: Bases de datos transaccionales, archivos de textos, hipertextos, hojas de cálculo, informes periódicos, entre otros.

- **Gestor de Carga**

  Son los procedimientos y técnicas para la integración de los datos; conocidos como extracción, transformación y carga (ETL).

  Extracción, permite la exploración de las diversas fuentes de datos disponibles y se extrae el dato seleccionado teniendo en cuenta su relevancia y las necesidades de los usuarios. Una vez que los datos son seleccionados y extraídos, se guardan en un almacenamiento intermedio, lo cual permite:

  - Manipular los datos sin interferir o suspender los procesos transaccionales en el DW.
  - No depender de la disponibilidad de los sistemas transaccionales.

- Almacenar y gestionar los metadatos que se generarán en los procesos extracción, trasformación y carga de datos.
- Facilitar la integración de las diversas fuentes de datos internas y externas.

Transformación de los datos, proceso que detecta las inconsistencias y los convierte en un conjunto de datos compatibles y congruentes, para que puedan ser cargados en el DW. Esto ocurre debido a que pueden existir diferentes fuentes de información, y hay necesidad de estandarizarlos e integrarlos en una única para ser ingresados al DW.

Carga, proceso que se encarga de la carga inicial de los datos al DW, así como de la actualización o mantenimiento periódico de los datos.

La carga inicial, resulta ser una tarea pesada y lenta, ya que se deben cargarse todos los registros que han sido generados. Mientras que la actualización, resulta más ligera y consiste en ir añadiendo al depósito solo aquellos datos nuevos que fueron generados desde la última actualización.

- **Gestor del Data Warehouse**

Es el almacén de datos que contiene los datos procesados de toda la organización, estos datos pueden ser incluso redundantes, sus características son:

- Se forma al integrar un Sistema de Gestor de Base de Datos, con software y aplicaciones dedicadas.
- Almacena los datos de forma multidimensional (tablas de hechos y tablas de dimensiones).
- Gestiona las diferentes estructuras de datos que se construyan o describan sobre el DW.
- Gestiona y mantiene el metadato.
- Transformar e integrar los datos fuentes y el almacenamiento intermedio en un modelo adecuado para la toma de decisiones.

- **Gestor de Consulta**

Son las herramientas que facilitan el acceso a los datos, mediante las cuales el usuario puede realizar la recuperación y el análisis de los datos.

2.3 Modelos de Almacenamiento

El modo de almacenamiento de los datos en un DW impacta en forma directa en el rendimiento de las consultas y el procesamiento, por tanto es necesario tener un especial cuidado en su elección entre los principales modelos: MOLAP, ROLAP y HOLAP.

- **MOLAP (Multidimensional - Online Analytical Processing)**

  Este modelo utiliza las bases de datos multidimensionales (BDMD) para proporcionar el análisis, permitiendo la visualización de la información en varias dimensiones. Un sistema MOLAP usa una base de datos propietaria multidimensional, en la que la información se almacena multidimensionalmente, para ser visualizada en varias dimensiones de análisis.

  MOLAP utiliza una arquitectura de **dos niveles**: la bases de datos multidimensionales y el motor analítico. La base de datos multidimensional es la encargada del manejo, acceso y obtención del dato.

  La información procedente de los sistemas operacionales, se carga en el sistema MOLAP, mediante una serie de rutinas por lotes. Una vez cargado el dato elemental en la Base de Datos multidimensional (MDDB), se realizan una serie de cálculos por lotes, para calcular los datos agregados, a través de las dimensiones de negocio, rellenando la estructura MDDB.

  Tras rellenar esta estructura, se generan unos índices y algoritmos de tablas hash para mejorar los tiempos de accesos a las consultas. Una vez que el proceso de compilación se ha acabado, la MDDB está lista para su uso. Los usuarios solicitan informes a través de la interfase, y la lógica de aplicación de la MDDB obtiene el dato.

  La arquitectura MOLAP requiere unos cálculos intensivos de compilación. Lee de datos precompilados, y tiene capacidades limitadas de crear agregaciones dinámicamente o de hallar ratios que no se hayan precalculados y almacenados previamente.

En Tabla 2 se muestra las ventajas y desventajas del modelo MOLAP.

Ventajas	Desventajas
• Mayor performance en el procesamiento de queries. • Poco tiempo de cálculo realizado en el momento. • Puede escribirse sobre la base de datos. • Posibilita hacer cálculos más complicados	• Tamaño limitado para la arquitectura del cubo. • No puede acceder a datos que no están en el cubo. • No puede explotar el paralelismo en las bases de datos.

Tabla 2: Ventajas y desventajas del modelo MOLAP

Fuente: Basado en Ibarra, María (2006)

- **ROLAP (Relacional - Online Analytical Processing)**

El sistema ROLAP utiliza una arquitectura de **tres niveles:** (i) El nivel de base de datos usa bases de datos relacionales para el manejo, acceso y obtención del dato, (ii) El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios y (iii) El motor ROLAP se integra con niveles de presentación, a través de los cuáles los usuarios realizan los análisis OLAP.

Después de que el modelo de datos para el DW se ha definido, los datos se cargan desde el sistema operacional. Se ejecutan rutinas de bases de datos para agregar el dato, si así es requerido por el modelo de datos. Se crean entonces los índices para optimizar los tiempos de acceso a las consultas.

Lo que hace con ROLAP es reorganizar los datos utilizando tablas o vistas agregadas que le permitan al motor ROLAP un mejor acceso. Este tipo de organización lógica suele lograrse mediante procesos ETL.

En Tabla 3 se muestra las ventajas y desventajas del modelo ROLAP.

Ventajas	Desventajas
• Uso total de la seguridad e integridad de la base de datos. • Escalable para grandes volúmenes. • Los datos pueden ser compartidos con diversas aplicaciones. • Los datos y estructura son más dinámicos. • Aprovechamiento de la infraestructura disponible.	• Queries más lentos. • Los cálculos están limitados a las funciones de las bases de datos.

Tabla 3: Ventajas y desventajas del modelo ROLAP

Fuente: Basado en Ibarra, María (2006)

- **HOLAP (Hybrid OLAP)**

  Es una combinación de los modelos ROLAP y MOLAP, para brindar una solución con las mejores características de ambas: desempeño superior y gran escalabilidad

  En este modelo, los datos agregados y precalculados se almacenan en estructuras multidimensionales (MOLAP) y los de menor nivel de detalle en el relacional (ROLAP). Para implementar este tipo de modelo se requiere un minucioso trabajo de análisis para identificar cada tipo de dato.

2.4 Inteligencia de Negocios

La gestión de la información en las organizaciones es, hoy en día, una herramienta clave para poder sobrevivir en un mercado cambiante, dinámico y global. Aprender a competir con esta información es fundamental para la toma de decisiones, el

crecimiento y la gestión de la organización. La Inteligencia de Negocios o Business Intelligence (BI) nos acerca a los sistemas de información que nos ayudan a la toma de decisiones en la organización.

2.4.1 Definición de Inteligencia de Negocios[9]

Inteligencia de Negocios es el conjunto de herramientas y estrategias cuyo objetivo básico es apoyar de forma sostenible y continua a las organizaciones para mejorar su competitividad, facilitando la información necesaria para la toma de decisiones. El primero que acuñó el término fue Howard Dresner (1989), cuando era consultor de Gartner, popularizó Business Intelligence o BI como un término paraguas para describir un conjunto de conceptos y métodos que mejoraran la toma de decisiones, utilizando información sobre qué había sucedido (hechos).

Mediante el uso de tecnologías y las metodologías de BI pretendemos convertir datos en información y a partir de la información ser capaces de descubrir conocimiento.

Para definir BI partiremos de la definición del glosario de términos de Gartner[10]:

"*BI es un* ***proceso interactivo*** *para* ***explorar*** *y* ***analizar información estructurada*** *sobre un* ***área*** *(normalmente almacenada en un DW), para descubrir tendencias o patrones, a partir de los cuales derivar ideas y extraer* ***conclusiones****. El proceso de Business Intelligence incluye la comunicación de los descubrimientos y efectuar los cambios. Las* ***áreas*** *incluyen clientes, proveedores, productos, servicios y competidores*."

2.4.2 Componentes de Inteligencia de Negocios

Los componentes de BI se muestran en el Figura 6 que a continuación se muestra.

---

[9] Business Intelligence: Competir con Información, Josep Lluis Cano, 2007

[10] Glosario de Gartner, 2006.

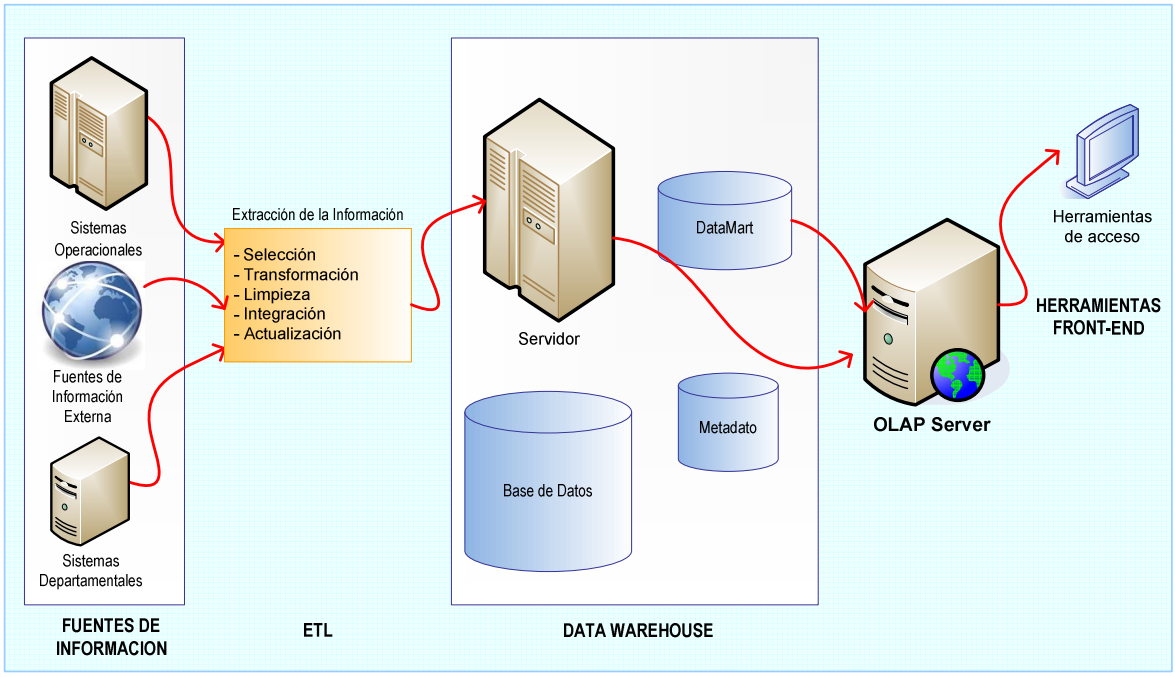

Figura 6: Componentes de la Inteligencia de Negocios

Fuente: Cano, Josep Lluis (2007)

- Fuentes de información, que servirán para alimentar de información el DW.
- Proceso de extracción y transformación de los datos para su carga en el DW.
- El propio DW o almacén de datos, con el Metadato o Diccionario de datos.
- El motor OLAP, provee la capacidad de cálculo, consultas, funciones de planeamiento, pronóstico y análisis de escenarios en grandes volúmenes de datos.
- Las herramientas Front End o Herramientas de visualización, permite la exploración de datos del DW. Estas herramientas constituyen el nexo entre el depósito de datos y el usuario.

2.5 Censos Nacionales

2.5.1 Definición

El Censo es el conjunto de operaciones destinadas a recopilar, procesar, evaluar y publicar datos de todas las unidades de un universo en un momento determinado. Etimológicamente, la palabra censo proviene del latín CENSERE que significa contar.

El censo es la única fuente de carácter universal que se ejecuta periódicamente para saber qué tenemos y qué nos falta. También se dice que es una fotografía que revela una imagen en un momento determinado. La información que se recoge en un censo, está en función a las necesidades de planificación o para la ejecución de políticas y programas de desarrollo a nivel nacional, regional y local.

La información del censo está referida a la totalidad de la población en estudio y los datos producidos pueden ser utilizados para realizar análisis pormenorizados tanto a escalas geográficas muy pequeñas como para grupos específicos.

La información censal, usualmente son recogidas a intervalos regulares de tiempo y nos permiten conocer las tendencias entre censo y censo.

2.5.2 Tipos de Censos

De acuerdo a la naturaleza de la información que se recopilan, los censos más importantes que ejecuta el INEI son: de población, de vivienda, agropecuario y económico.

- **Censo de Población**

  El Censo de Población es el conjunto de operaciones destinadas a recoger, recopilar, evaluar, analizar y publicar datos demográficos, económicos, sociales y culturales, de todos los habitantes de un país, en un momento o período indicado. Las características demográficas se refieren al número de habitantes, su distribución por sexo y edad, natalidad, mortalidad y migración. Las económicas comprenden la profesión u oficio, la condición y categoría de ocupación, entre otras. Las culturales y sociales están referidas al nivel de instrucción, asistencia escolar, idioma, etc.

  Considerando el lugar de residencia de la población, los censos pueden ser:

  Censo de Hecho, cuando se empadrona a las personas en el lugar donde se encontraron la noche anterior al "Día del Censo".

De Derecho, cuando se empadrona a las personas en el lugar de su residencia habitual.

- **Censo de Vivienda**

  El Censo de Vivienda es un conjunto de operaciones destinadas a recoger, recopilar, evaluar, analizar y publicar datos estadísticos correspondientes a todas las viviendas y sus ocupantes, en un momento o período indicado. Este censo nos informa sobre el número de habitantes de la vivienda, el material con el que está construida, el régimen de tenencia, los servicios que dispone, tipo de alumbrado, entre otras características.

- **Censo Agropecuario**

  Es la fuente de información estadística más importantes sobre la estructura agropecuaria del país, dado que proporciona datos actualizados para el mejor conocimiento de la estructura y base productiva agropecuaria del país que permitirá la formulación de planes de desarrollo y la adopción de políticas que coadyuven a mejorar el nivel de vida de la población vinculada a las actividades agropecuarias.

- **Censo Económico**

  Constituye la fuente de información más completa sobre la estructura económica del país en un momento determinado. Proporciona información clasificada por actividad económica y por región geográfica sobre las características de las unidades económicas del país, como: el número de empresas y establecimientos, la cantidad de personal ocupado, las remuneraciones, ingresos, egresos, valor agregado y activos fijos.

  El diagnostico que se obtiene con este censo sobre la estructura económica y productiva del país permite evaluar las tendencias de la economía y diseñar medidas de política económica que favorezcan el desarrollo.

2.5.3 Importancia de los Censos Nacionales de Población y Vivienda

Los censos son importantes porque constituyen la única fuente de información que proporciona datos al menor nivel de desagregación geográfica de todo el país. Por lo cual, es posible la elaboración de programas de desarrollo social y económico hasta el mínimo nivel político administrativo. Así mismo, permite conocer variables derivadas como la mortalidad, fecundidad y migraciones, y posibilite realizar estudios para disminuir la pobreza y focalizar la inversión social.

Igualmente, proporcionan el marco muestral adecuado para ejecutar estudios detallados de interés social, demográfico y económico, entre otros. Por lo general estos censos se ejecutan cada 10 años.

También las municipalidades demandan la ejecución del Censo de Población y Vivienda en sus áreas jurisdiccionales, a fin de conocer la magnitud de su población que permitirá a los alcaldes distritales justificar mayores requerimientos presupuestales ante el Tesoro Público, para la ampliación de su infraestructura distrital, atender la demanda de servicios básicos para un volumen población real y participación en programas sociales.

En resumen la información de los censos nacionales de población y vivienda se utiliza para:

- Conocer la distribución de la población en el territorio nacional.
- Mejorar la planificación y ejecución de los programas sociales y de asistencia alimenticia.
- Ubicar nuevos centros de salud, escuelas y otros servicios esenciales.
- Identificar las zonas con dificultad de acceso a los servicios de agua potable, saneamiento y electricidad.
- Medir el nivel educativo de la población.
- Conocer las tasas de natalidad y mortalidad en todo el país y en áreas específicas.
- Lograr que las empresas privadas definan mejor su mercado.
- Conocer las características de las viviendas, como el tipo de material utilizado en la construcción y servicios básicos que dispone.

- Identificar el nivel de bienestar de los hogares, mediante la tenencia de artefactos electrodomésticos y servicios de Internet, cable, etc.

2.5.4 Etapas de los Censos Nacionales de Población y Vivienda

Los censos nacionales de población y vivienda se ejecutarán teniendo en cuenta las etapas y actividades que se muestran en el Figura 7.

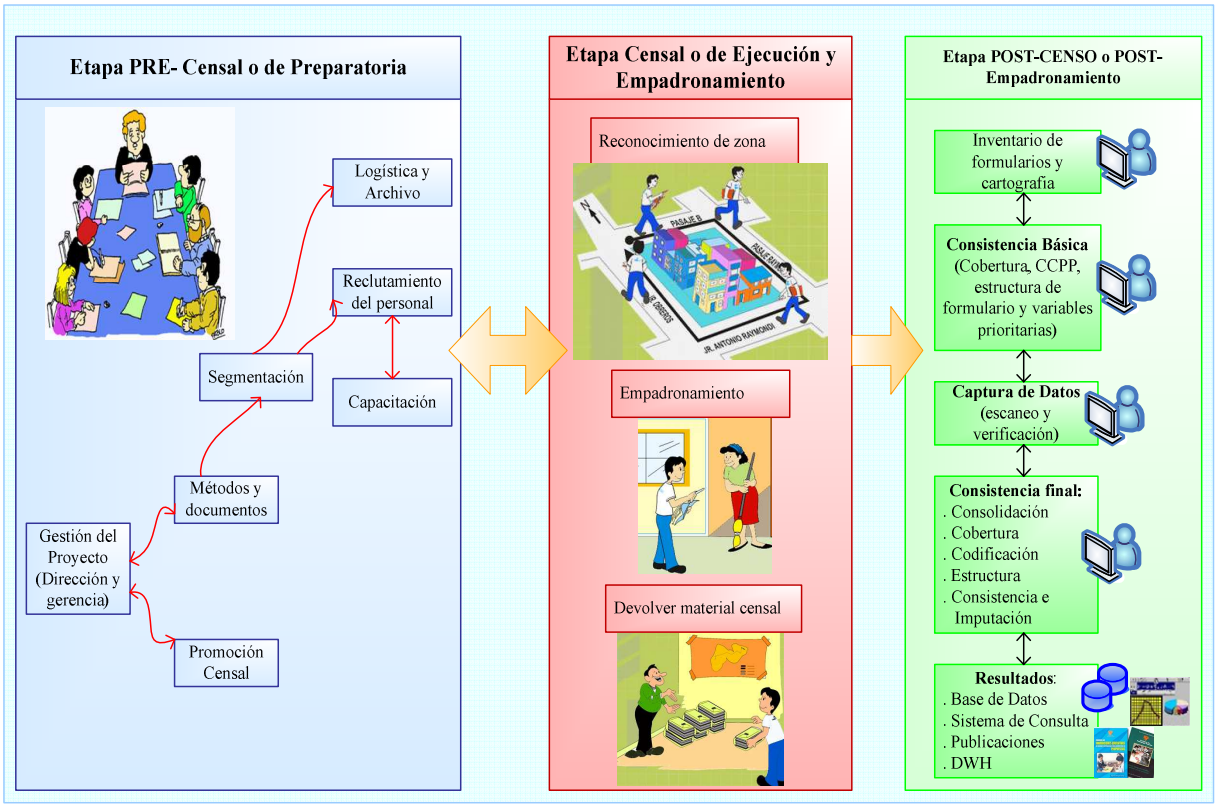

Figura 7: Etapas de los Censos Nacionales

Fuente: Elaboración Propia

Donde:

1) Etapa PRE- Censal o de Preparatoria

En esta etapa se realizarán las actividades censales siguientes:

a) Gestión del Proyecto

Consiste en la planificación de cada actividad censal, su normatividad técnica y administrativa y el control de su ejecución.

b) Métodos y Documentos

Establece los procedimientos técnicos metodológicos y asegurar la uniformidad conceptual para estandarizar la ejecución de los procesos censales. Aquí es donde se desarrolla la temática censal, la elaboración de los diseños muestrales para el control de calidad de las tareas ejecutadas, el diseño de la cédula censal, los manuales, documentos auxiliares y demás documentos.

c) Segmentación

Es la ejecución de un conjunto de diversas tareas de campo y gabinete, dirigidas a dividir el territorio de cada distrito del país en áreas de empadronamiento urbano y rural. Estas áreas deben tener límites claramente definidos y de fácil identificación en el terreno, los mismos que están representados en los documentos cartográficos.

También se preparan la documentación cartográfica, material para la capacitación del personal, así como la elaboración de legajos para el empadronamiento y elaboración de informes. Todo ello con el propósito de evitar omisiones o duplicaciones de las viviendas y/o personas, durante el empadronamiento.

d) Capacitación para el Empadronamiento

En los censos nacionales, la capacitación es uno de los procesos más importantes por su repercusión e impacto en la calidad de los resultados censales. Por lo que demanda una adecuada planificación y una selección de metodologías que apoye la enseñanza y facilite el aprendizaje.

En atención a la naturaleza de las funciones del personal para el empadronamiento, se requiere organizar y diseñar programas de capacitación para dos grupos de funcionarios censales: Del personal que tiene la responsabilidad de la organización, conducción, supervisión y control, y del personal de campo encargado de ejecutar el empadronamiento.

e) Promoción Censal

La promoción censal para los Censos Nacionales, comprende un conjunto de tareas dirigidas a establecer diversos canales de comunicación orientados a desarrollar una "cultura de censos" en la opinión pública.

Tiene por finalidad, dar a conocer a la población, de manera gradual y sistemática, las características principales de los censos nacionales, con el objeto de interesar y motivar su participación, creando un clima favorable para la ejecución de cada una de las etapas censales.

Así mismo, se realizan un conjunto de acciones para lograr el apoyo de autoridades políticas y administrativas del gobierno central, regional y local; así como de los organismos internacionales, personalidades, empresas privadas, docentes y estudiantes, que coadyuven al desarrollo de las actividades censales.

f) Logística y Archivo

Es la ejecución de un conjunto de tareas para distribuir oportunamente, desde la sede central a las oficinas departamentales, los documentos y material necesarios para la ejecución de las actividades censales. Así mismo, comprende la recepción en la sede departamental, de los documentos diligenciados en su jurisdicción para su inventario, ordenamiento, clasificación y archivo.

g) Reclutamiento del personal para el empadronamiento

Son las actividades destinadas a disponer de manera simultánea a nivel nacional del personal necesario que serán responsables de la ejecución del empadronamiento censal.

El proceso de reclutamiento requiere ser muy cuidadoso, preciso y dinámico. En primer lugar, por la calidad del personal a reclutar; segundo, por la cantidad de personas necesarias; y, tercero, por los tiempos en la ejecución de los censos.

2) Etapa Censal o de Ejecución y Empadronamiento

Esta actividad consiste en la recopilación de información sobre la población y las viviendas existentes, mediante entrevista directa del empadronador a los integrantes de cada hogar usando la cédula censal.

3) Etapa de Post- Empadronamiento

Esta etapa se refiere al procesamiento de los datos, a la preparación, edición, impresión y difusión de los resultados censales, a través de ediciones populares, cartillas de divulgación y el uso de medios informáticos.

Capítulo III: ESTADO DEL ARTE METODOLOGICO

3.1 Aplicaciones

Las referencias de trabajos o aplicaciones similares a la presente tesina son los siguientes:

3.1.1 Instituto Nacional de Estadística, Geografía e Informática (INEGI) de México: Data Warehouse para la prestación del Servicio Público de Información Estadística [11]

El INEGI es la entidad que tiene la responsabilidad de coordinar los Sistemas Nacionales de Estadística y de Información Geográfica de México, además de promover y orientar el desarrollo informático en el país.

Esta entidad disponía de mucha información de diversos proyectos estadísticos (censos, encuestas y registros administrativos) y tenían necesidad de contar con una única fuente de datos definitiva donde la información estuviese integrada, que la interpretación de la información esté asistida por metadatos, que la información sea independiente de las aplicaciones y el aprovechamiento sea por todos los usuarios internos y externos al INEGI.

Para tal fin, creó el proyecto "DW para la Prestación del Servicio Público de Información Estadística", que consistió en el análisis, diseño, desarrollo,

[11] Data Warehouse para la Prestación del Servicio Público, 2003

implementación, mantenimiento y mejora continua del repositorio de información estadística del país, que proporcione una herramienta para consulta y análisis a los usuarios.

De este modo, el almacén de datos o DW estadístico, se incorporó en los procesos de producción de información estadística, y se convirtió en la fuente de datos única para la presentación de los resultados, que ahora no se limita a la generación de publicaciones sino que además es la fuente de información para sistemas de consultas internas, desarrollo de productos digitales para el sitio de INEGI en Internet.

3.1.2 Instituto Nacional de Estadística (INE) de España: Resultados Definitivos de los Censos de Población y Viviendas 1991 (INEbase)

El INE de España es un organismo autónomo que en la última década ha incorporado de manera intensa el uso de las nuevas tecnologías en el desarrollo estadístico. Para los Censos de Población y Vivienda de 2001, simplificaron el procesamiento de datos y utilizó por primera vez un DW para difundir los resultados, el cual ha tenido una excepcional acogida entre los usuarios.[12]

El proyecto consideró tres etapas[13]:

- **Modelado multidimensional,** se identificaron los colectivos (Personas, Edificios, Viviendas, Hogares) de interés en base a las variables o dimensiones que tienen en común. También se evaluó el impacto del amplio número de dimensiones y de las restricciones para el almacenamiento físico y lógico usando modelos multidimensionales.

- **Trabajo con los datos**, definición de si son tablas o cubos los que se van a construir, luego la construcción de los diferentes niveles de agregación previa definición de cuántos y cuáles son los objetos presumarizados a crear. Para lo

[12] Sociedad de Estadística e Investigación Operativa, 2006
[13] El INE y su producción estadística, 2011

cual son necesarios los procesos de Extracción, Transformación y Carga de los Datos.

- **Diseño de Interfaz,** a fin de mostrar una aplicación de fácil uso para todo tipo de usuarios.

Este DW permitió a los usuarios diseñar sus propias tablas, estableciendo las condiciones que debe tener la información, aplicando en todo momento la restricción que impone la confidencialidad estadística.

Este almacén de datos tiene las funcionalidades siguientes: [14]

a) Lista de tablas predefinidas que se consideran de especial interés, para los usuarios que no tengan una idea previa totalmente formada de lo que necesitan y los censos pueden ofrecerle.

b) Posibilidad de 'crear tablas', que permite a cada usuario definir exactamente el cruce que necesita. Esta funcionalidad es la que realmente hace un cambio bastante notable con respecto a la forma de consulta de censos anteriores, en los que a menudo los usuarios debían realizar peticiones a medida porque las tablas publicadas no se ajustaban exactamente a sus necesidades.

c) La información solicitada se responderá on-line o será enviada a un proceso batch si la consulta es muy pesada, pero cuyo tiempo de respuesta será de menos de 24 horas. Sólo las peticiones que, por su volumen o que requieran realizar cálculos laboriosos con los microdatos tardarán algo más; en cualquier caso, mucho menos que con el sistema anterior.

d) A partir de la consulta inicial, la misma interfaz permite ir afinando interactivamente la petición, añadiendo nuevas variables, filtros, cambiando la unidad de medida, etc, hasta llegar a la tabla realmente requerida. El cual puede ser rotada, representada con gráficos o mapas, o exportada a formatos estándar para su posterior manipulación.

[14] La aportación del Censo de Viviendas 2001 de España

3.2 Metodologías para la Construcción de Data Warehouse[15]

En cuanto a su construcción, a la hora de abordar un DW no hay una única metodología para el diseño, sino que dependiendo del contexto en el que se encuentre la organización y los objetivos que persiga se puede emplear una u otra metodología. Estas diferentes metodologías se pueden englobar dentro de dos grandes bloques: top-down y bottom-up que corresponden a las metodologías propuestas por Bill Inmon (Enfoque CIF -Fabrica de Información Corporativa) y Ralph Kimball (Enfoque Dimensional) respectivamente. También existen otras metodologías alternativas como la Hefesto y The SAS Rapid DW que son adaptaciones de las anteriores y metodologías hibridas que son la combinación de las mismas como Hybrid DW, Federated DW y Barquin del Instituto de DW.

3.2.1 Metodología Bill Inmon

Bill Inmon considerado como el padre del DW, formaliza el concepto en 1992 con su libro "Building the Data Warehouse", el cual se constituyó en la primera guía para construir un DW y donde a pautas muy concretas, las cuales actualmente siguen vigentes.

En su enfoque Corporate Information Factory (CIF) mostrado en la Figura 8, propone una estrategia corporativa para la construcción de DW, donde plantea transferir la información de los diferentes Sistemas Transaccionales (OLTP) de las organizaciones a un lugar centralizado donde los datos puedan ser utilizados para el análisis.

Considera que la información debe estar lo más detallado posible. Los DW departamentales o Data Marts son tratados como subconjuntos del DW Corporativo, que son construidos para cubrir las necesidades individuales de análisis de cada departamento, y siempre a partir del DW Central.

[15] Ampliación de conceptos del Modelado Dimensional. Roberto Espinosa, 2010

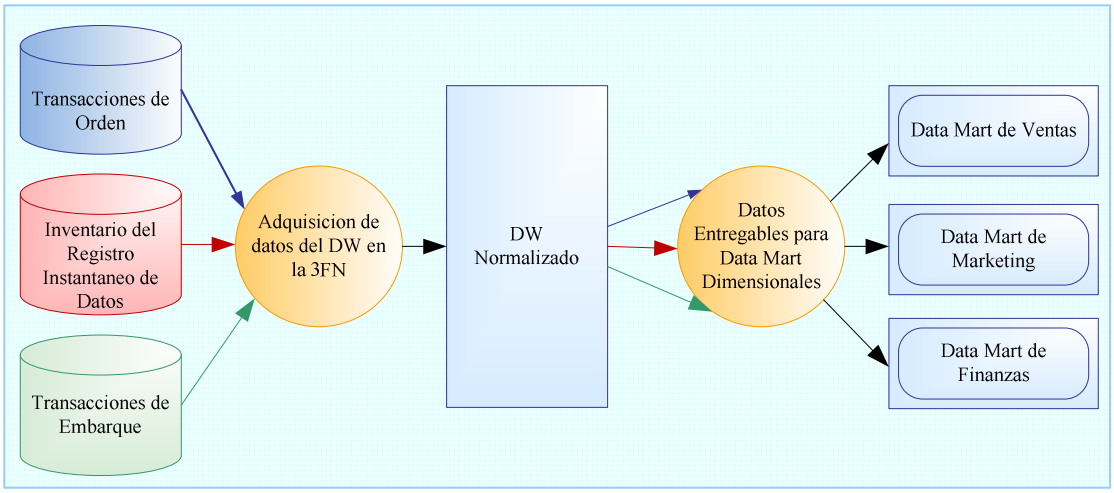

Figura 8: Metodología Inmon - Corporate Information Factory (CIF)

Fuente: Espinosa, Roberto (2010)

El enfoque de Inmon, referenciada normalmente como Top-Down, propone definir un DW Corporativo; y a partir de él ir construyendo los modelos de análisis para los distintos niveles y departamentos de la organización.

Por tanto, los datos son extraídos de los sistemas operacionales mediante procesos ETL y cargados en las áreas de stage, donde son validados y consolidados en el DW corporativo, además están los llamados metadatos que documentan de una forma clara y precisa el contenido del DW. Una vez implementado el DW; existen procesos de los Data Mart departamentales que obtienen la información de él con las consiguientes transformaciones, organizan los datos en las estructuras particulares requeridas por cada uno de ellos y refrescan su contenido.

La aplicación de la metodología de Inmon, obliga a tener este enfoque global, por lo que es más difícil para el desarrollo en un proyecto sencillo, pues se debe abordar el "todo" a partir del cual se irá al "detalle".

3.2.2 Metodología de Ralph Kimball

Ralph Kimball considerado como un "Gurú" del DW, en su libro "The Data Warehouse Lifecycle Toolkit" del año 1996, propone la metodología MD (Modelo Dimensional),

que se ha convertido en el estándar de facto para la construcción de DW como soporte a las decisiones empresariales.

Según este enfoque el DW es un conglomerado de todos los Data Marts dentro de una empresa; siendo una copia de los datos transaccionales estructurados de una forma especial para el análisis, de acuerdo al Modelo Dimensional (no normalizado) que incluye las dimensiones de análisis y sus atributos, así como los diferentes hechos de negocio que se quieren analizar. De este modo tenemos tablas para representar las dimensiones y por otro las tablas de hechos (facts tables).

Bajo este enfoque, los diferentes Data Marts están relacionados entre sí, por la llamada bus structure, que contiene los elementos anteriormente citados a través de las dimensiones conformadas, que permite que los usuarios puedan realizar querys sobre los diferentes Data Marts, pues este bus contiene los elementos en común que las integran. Una dimensión conformada puede ser, por ejemplo, la dimensión cliente, que incluye todos los atributos o elementos de análisis referentes a los clientes y que puede ser compartida por diferentes Data Marts (ventas, pedidos, gestión de cobros, etc).

Este enfoque conocido como **Bottom-up,** hace más flexible y sencillo de implementar, pues podemos construir un Data Mart como primer elemento del sistema de análisis, y luego ir añadiendo otros que comparten las dimensiones ya definidas o incluyen otras nuevas. En este sistema, los procesos ETL extraen la información de los sistemas operacionales y los procesan igualmente en el área stage, realizando posteriormente el llenado de cada uno de los Data Mart de una forma individual, aunque siempre respetando la estandarización de las dimensiones.

Para el enfoque de Kimball, el desarrollo de un proyecto DW debe considerar:

- Selección del proceso de negocio
- Definición de la granularidad de la información
- Elección de las dimensiones de análisis
- Identificación de los hechos o métricas.

En Figura 9 se muestra las etapas del enfoque de Ralph Kimball.

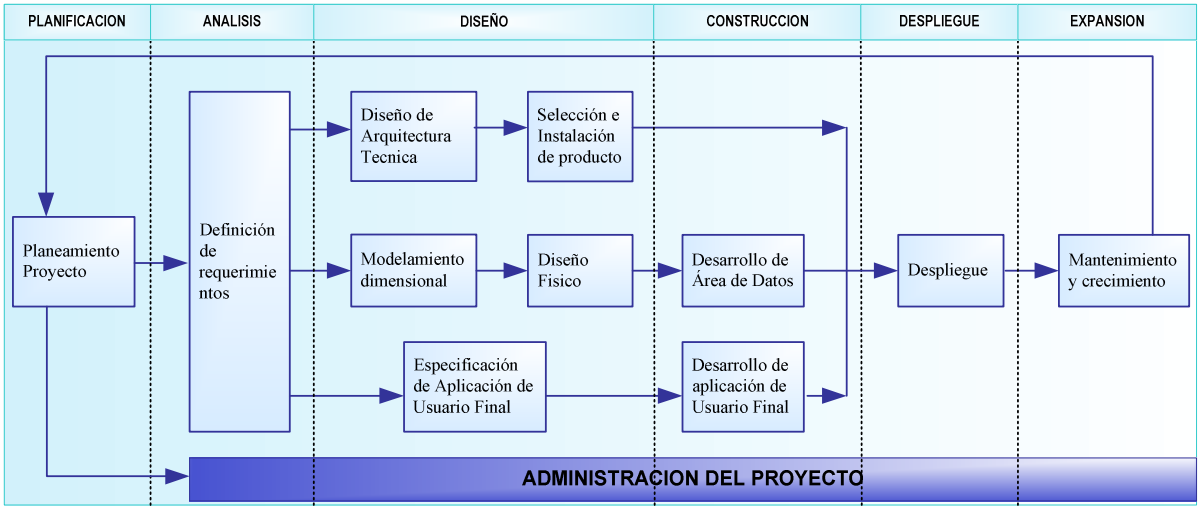

Figura 9: Enfoque Kimball de Construcción del Data Warehouse

Fuente: asado en Córdova Rodriguez, Diana (2011)

Donde:

1) **Planificación: Planificación del Proyecto**

La planificación del proyecto, busca identificar la definición y el alcance del proyecto de DW, incluyendo las justificaciones del negocio y las evaluaciones de factibilidad, las que se basan en los requerimientos del negocio y no en fechas límites. También establece personal (los usuarios, gerente del proyecto, equipo del proyecto), desarrollo del plan del proyecto, el seguimiento y el monitoreo.

Esta etapa se concentra sobre la definición del proyecto y según sentencia Kimball: "*Antes de comenzar un proyecto de DW o Data Mart, hay que estar seguro si existe la demanda y de dónde proviene. Si no se tiene un usuario sólido, posponga el proyecto*".

2) **Análisis: Definición de Requerimientos**

Un factor determinante en el éxito de un proyecto de DW, es la interpretación correcta de los diferentes niveles de requerimientos expresados por los distintos grupos de usuarios.

La técnica utilizada para conocer los requerimientos de los analistas del negocio difiere de los enfoques tradicionales guiados por los datos. Los diseñadores de los DW deben entender los factores claves que guían el negocio para determinar efectivamente los requerimientos y traducirlos en consideraciones de diseño apropiadas.

Los usuarios finales y sus requerimientos impactan siempre en la implementación de un DW. Según la perspectiva de Kimball, los requerimientos del negocio se posicionan en el centro del "universo del Data Warehouse", por tanto los requerimientos del negocio deben determinar el alcance del DW (qué datos debe contener, cómo deben estar organizados, cada cuánto tiempo debe actualizarse, quiénes y desde dónde accederán, etc.).

3) Diseño:

a) Diseño de la Arquitectura Técnica

Las soluciones de los DW requieren la integración de numerosas tecnologías y para poder establecer un diseño de la arquitectura técnica es necesario tener en cuenta tres factores claves: los requerimientos del negocio, los actuales entornos técnicos y las directrices técnicas y estratégicas futuras planificadas por la empresa.

Algunos equipos de trabajo no comprenden las ventajas de una arquitectura y tienen la sensación de que las tareas son demasiado ambiguas, por lo que asumen su diseño como una distracción y un obstáculo para el desarrollo del DW, así que optan por omitir el diseño de la arquitectura. Sin embargo, hay otros equipos de trabajo que dedican un tiempo demasiado grande para el diseño arquitectónico. Ralph Kimball recomienda no irse a ninguno de los dos extremos.

b) Selección e Instalación de Productos

Utilizando el diseño de arquitectura técnica como base, es necesario evaluar y seleccionar los componentes específicos de la arquitectura, como la plataforma de hardware, el motor de base de datos, la herramienta de ETL, las herramientas de acceso, etc. Una vez evaluados y seleccionados los componentes, se procede con la instalación y prueba de los mismos en un ambiente integrado de DW.

c) **Modelado Dimensional**

La definición de los requerimientos del negocio determina los datos necesarios para cumplir los requerimientos analíticos de los usuarios. Diseñar los modelos de datos para soportar estos análisis, requiere un enfoque diferente al usado en los sistemas operacionales. Básicamente, se comienza con una matriz donde se determina la dimensionalidad de cada indicador y luego se especifican los diferentes grados de detalle dentro de cada concepto del negocio, así como la granularidad de cada indicador y las diferentes jerarquías que dan forma al modelo dimensional del negocio o mapa dimensional.

d) **Diseño Físico**

El diseño físico de la base de datos se focaliza sobre la selección de las estructuras necesarias para soportar el diseño lógico. Un elemento principal de este proceso es la definición de estándares del entorno de la base de datos. En esta etapa también se determinan la indexación y las estrategias de particionamiento.

e) **Especificación de Aplicación de Usuario Final**

No todos los usuarios del DW necesitan el mismo nivel de análisis. Es por ello que en esta etapa se identifican los roles o perfiles de usuarios para los diferentes tipos de aplicaciones necesarias en base al alcance de los perfiles detectados (gerencial, analista del negocio, vendedor, etc.)

4) **Construcción:**

a) **Diseño y desarrollo de la presentación de datos**

Entre las actividades más importantes de esta fase son los Procesos ETL.

Extracción. Son los procesos requeridos para obtener los datos que permitirán efectuar la carga del Modelo Físico diseñado

Transformación. Procesos para convertir o recodificar los datos fuente a fin de poder efectuar la carga efectiva del Modelo Físico

Carga. Son los procesos requeridos para poblar el DW.

Todas estas tareas son altamente críticas pues tienen que ver con la materia prima del DW: los datos. La desconfianza y pérdida de credibilidad del DW provocará efectos inmediatos e inevitables si el usuario se encuentra con información inconsistente. Es por ello que la calidad de los datos es un factor determinante en el éxito de un proyecto de DW. Es en esta etapa donde deben sanearse todos los inconvenientes relacionados con la calidad de los datos fuente. Para cumplir con estas premisas es necesario tener en cuenta ciertos parámetros a la hora de desarrollar las tablas de dimensión y la tabla de hechos.

b) Desarrollo de Aplicación de Usuario Final

Involucra configuraciones de los metadatos y construcción de la aplicación de acuerdo a especificaciones.

Los usuarios acceden al DW por medio de herramientas de productividad basadas en GUI (Graphical User Interface). De hecho existen multitud de estas herramientas con las que proveer a los usuarios, estas herramientas pueden incluir software de consultas, generadores de reportes, procesamiento analítico en línea o herramientas de Datamining dependiendo de los tipos de usuarios y sus requerimientos particulares. Sin embargo, una sola herramienta puede no satisfacer todos los requerimientos, por lo que quizás sea necesario la integración de herramientas hechas bajo demanda expresa de los usuarios para satisfacer sus necesidades de consulta sobre el DW.

5) Despliegue:

Representa la convergencia de la tecnología, los datos y las aplicaciones accesibles para el usuario del negocio. Hay varios factores extras que aseguran el correcto funcionamiento de todos estos elementos, entre ellos se encuentran la capacitación, el soporte técnico, la comunicación y las estrategias de feed back. Es decir que la información contenida en el DW debe estar en términos y lenguajes que sean comprendidos por los usuarios, quienes no son técnicos.

6) **Mantenimiento y Crecimiento:**

Como se remarca siempre, la creación de un DW es un proceso (de etapas bien definidas, con comienzo y fin, pero de naturaleza espiral) que acompaña a la evolución de la organización durante toda su historia. Se necesita continuar con las actualizaciones de forma constante para poder seguir la evolución de las metas por conseguir. Al contrario de los sistemas tradicionales, los cambios en el desarrollo deben ser vistos como signos de éxito. Es importante establecer las prioridades para poder manejar los nuevos requerimientos de los usuarios y de esa forma poder evolucionar y crecer. Una vez que se ha construido e implantado el DW se debe estar preparado para administrar el mantenimiento y crecimiento del mismo. Si bien las tareas pueden parecer similares a las tratadas en otras etapas del ciclo de vida, existe una diferencia clave: los usuarios están ahora accediendo al DW.

7) **Administración del Proyecto**

La administración del proyecto asegura que las actividades del ciclo de vida se lleven a cabo de manera sincronizada. La administración del proyecto acompaña todo el ciclo de vida. Entre sus actividades principales se encuentra el monitoreo del proyecto y el acoplamiento entre los requerimientos del negocio y las restricciones de los sistemas de información para poder manejar correctamente las expectativas en ambos sentidos.

3.2.3 Metodología Ramón Barquín

Ramón Barquín, co-fundador y primer presidente The Data Warehousing Institute, autor de la metodología que lleva su nombre, quien publicara "Planning and Designing the Data Warehouse" y "Building, Using and Managing the Data Warehouse", para la construcción de un DW propone un análisis de la estrategia, la arquitectura y la tecnología de manera integral, considerando al DW como un todo. Dicha metodología propone también una construcción incremental, por áreas temáticas de negocio. Los pasos propuestos por esta metodología se muestran en la Figura 10.

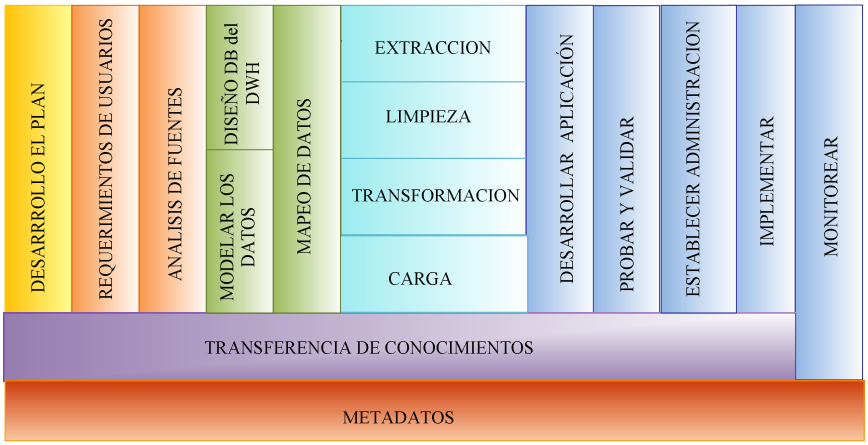

Figura 10: Metodología de Ramón Barquín para el desarrollo del Data Warehouse

Fuente: Barquin, Ramón (2009)

1) Desarrollar el Plan

El desarrollo del Plan, implica definir y establecer los objetivos a ser cumplidos a nivel de detalle, esta focalizado en la construcción del DW, de acuerdo a las prioridades dictadas por la estrategia.

Como entregable de esta etapa se tiene el Plan del Proyecto, el cual debe contener los pasos a ser seguidos, recursos que se requieran, personas y habilidades requeridas, tecnología a usar, los materiales, costos, cronogramas, riesgos y un plan de contingencias.

2) Requerimientos de los usuarios

Recolectar toda la información necesaria para desarrollar el proyecto, para ello se debe identificar a los usuarios y realizar el descubrimiento de la información utilizando las técnicas de recolección de requerimientos, tales como entrevistar a los usuarios, revisión de metadatos existentes y la información en uso.

Como entregable se tiene el documento de requerimientos, el cual contiene adicionalmente las funciones del negocio, la data utilizada, las aplicaciones, la tecnología y la organización de la empresa.

Las tareas más importantes a realizar son:

- Conocer la información.
- Crear un Comité de alto nivel del Data Warehouse.
- Crear un Comité de usuarios del Data Warehouse.
- Validar los requerimientos.

- Alinear con la visión del negocio, la estrategia del DW.
- Alinear con la arquitectura corporativa de TI.

3) Análisis de Fuentes

Identificar los sistemas fuentes, para lo cual se debe estudiar y entender la arquitectura TI, realizar un inventario de los sistemas transaccionales existentes, de los sistemas de análisis existentes, investigar fuentes potenciales del DW, explorar e investigar fuentes externas a la empresa, explorar los temas de calidad de datos y entender la administración de cambios de los sistemas fuentes.

El entregable de esta etapa es el documento de Sistemas Fuentes, el cual debe contener la lista de los sistemas fuentes, su formato, plataforma del sistema, motor de Base de Datos, tamaño, frecuencia de actualización, calidad de los datos, dueño de los datos.

4) Modelar los Datos

El objeto es obtener un modelo de los datos origen, para ello se determina si existen modelos de datos y procesos del negocio, si existe debe ser usado para el siguiente paso, de lo contrario se debe revisar y validar los procesos de negocio, ligar a los requerimientos de negocio, desarrollar relaciones entre las funciones del negocio y modelar los procesos de negocio. En el modelado de la data considerar la herramienta de modelado de datos, establecer estándares de nombres, determinar la granularidad y validar el modelo de datos

El entregable de esta etapa es el Modelo de Datos de las fuentes origen.

5) Diseñar la Base de Datos del Data Warehouse

Es la etapa en la cual se realiza el diseño del DW, el diseño debe estar alineado a los requerimientos del negocio. Se considera el desarrollo de un Modelo Conceptual, luego el Modelo Lógico y finalmente el desarrollo del Modelo físico, en el cual se tienen consideraciones de los volúmenes, características de paralelismo y estrategias de segmentación. También se tiene en cuenta las necesidades de considerar los datos derivados. Finalmente se generan los scripts de creación de la Base de Datos.

El entregable en esta etapa es la Base de Datos física creada.

6) **Mapeo de Datos**

El Mapeo de Datos establece las relaciones existentes entre los datos origen y los datos destino en el DW. Se debe determinar el rol del Staging Area, donde se realizaran las transformaciones de los datos. Considerar la creación de datos agregados, totales y subtotales, etc.

El entregable de esta etapa es la Matriz de Mapeo de Datos, el cual sirve como insumo para los procesos de ETL.

7) **Extracción de datos**

El objeto es extraer los datos requeridos y colocarlos en el Staging Area (o direccionarlos en el DW destino).

El entregable de esta etapa son los procesos de Extracción de Datos, el cual contiene rutinas para seleccionar los campos de las fuentes, los cuales pueden estar en diferentes formatos. En Figura 11 se muestra en forma grafica el proceso de extracción de datos.

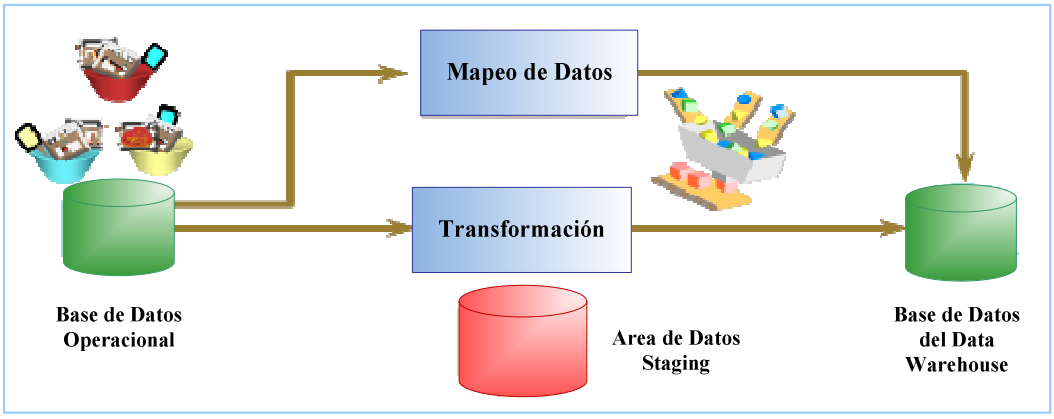

Figura 11: Procesos de extracción de datos

Fuente: Barquín, Ramón (2009)

8) **Limpieza de datos**

En esta etapa se considera la necesidad de limpieza, sincronización y estandarización de los datos. Se deben establecer métricas de calidad mínima, identificar fuentes de datos, medir la calidad de los datos, usar reglas de negocio para identificar inconsistencias, errores, incompatibilidad, o duplicidad en los datos. Se deben cargar solo los datos limpios, identificando y corrigiendo la

causa de los defectos, además programar limpiezas periódicas de las fuentes de datos.

9) **Transformación de los datos**

En esta etapa se desarrollan los procesos de transformación de datos, para lo cual se detallan y describen las derivaciones necesarias, sumarizaciones u otras operaciones. La transformación elimina anomalías de la data operacional.

Los entregables de esta etapa son los programas de transformación y la documentación respectiva.

10) **Cargar el Data Warehouse**

En esta etapa se realizan los procesos de carga, para lo cual se desarrolla el plan de carga, que debe considerar el tiempo de ejecución de los procesos de carga que deben ser realizados.

11) **Implementar el Metadato**

El metadato son los datos acerca de los datos y es tan importante como el propio dato del warehouse y debe ser generado en todos los pasos del proceso.

En general, un repositorio de Metadatos debe contener lo siguiente:

- Nombres de campos , definiciones y dominios
- Mapeo de los datos
- Tablas
- Índices
- Cronogramas de extracción, carga, etc.
- Criterios de selección
- Cálculos de los datos derivados
- Transformación de los datos

12) **Establecer los procesos de Administración**

En esta etapa se desarrolla un plan de operación y mantenimiento del DW. Para lo cual se establece un plan de administración de las operaciones de back-end, un plan de administración de las operaciones del metadato, un plan de

administración de las operaciones de acceso de los usuarios finales, un plan de administración del cambio y un plan de la administración del control de acceso y la seguridad.

Es importante considerar los temas relacionados a la seguridad del sistema, identificando los niveles de seguridad requeridos, los algoritmos de encriptación requeridos, los perfiles de acceso al sistema y las auditorias que se realizaran.

Los entregables son el plan de operación y mantenimiento y el plan de seguridad del sistema.

13) Desarrollar Aplicación

Esta es la etapa de la creación de la aplicación que accederá al DW y presentará la información a los usuarios, las aplicaciones deben estar alineadas a la visión del negocio y los requerimientos del usuario. Las aplicaciones se desarrollan por área de negocio. Se recomienda desarrollar prototipos de las pantallas del sistema y validar con los usuarios finales. Se debe tener en cuenta que si bien las aplicaciones OLAP (de análisis de procesos en línea) son las típicas explotaciones de un DW, también existen aplicaciones del tipo de Minería de Datos, Balanced Scorecard, Análisis de datos geoespacial.

El entregable de esta etapa es la aplicación de acceso al DW, la cual incluye la documentación del sistema.

14) Probar y validar el Data Warehouse

Se debe desarrollar un plan de prueba y validación, para ello se debe comprometer a los usuarios finales usando los "Data warehousing User's Advisory Group". En dicho plan se debe establecer claramente los parámetros y métricas de prueba.

El entregable de esta etapa es el sistema validado, y la documentación asociada es el plan de pruebas y los resultados obtenidos, en el cual se detallen ajustes y mejoras al sistema.

15) **Implementar**

Se desarrolla la implementación del sistema en el ambiente de producción, para ello se deben identificar a todos los usuarios, preparar la infraestructura de los usuarios y establecer entregables de usuarios. Se deberá preparar el ambiente de producción, desarrollar operaciones y la documentación de usuarios y publicar los documentos necesarios para su difusión.

Los entregables de esta etapa incluyen el sistema puesto en producción y la documentación de uso y mantenimiento del sistema.

16) **Transferencia de Conocimientos**

Se determina los requerimientos de entrenamiento (al staff y a los usuarios finales) necesarios para un uso adecuado del sistema, se identifica las características de los usuarios a quien se impartirá el entrenamiento y se desarrollará el plan y calendario del entrenamiento, además de diseñar el contenido del entrenamiento.

Los entregables de esta etapa es el personal entrenado en el uso del sistema y la documentación del entrenamiento.

17) **Monitoreo del Sistema**

Se hace seguimiento al uso del sistema, se identifica quienes son los principales usuarios, cuales son las consultas típicas que se realizan, que datos son los más consultados, que reportes son los preferidos, cuales son los tiempos de respuesta, cual es la calidad de los datos, etc. Esta etapa genera información necesaria para realizar los ajustes al sistema.

Los entregables de esta etapa incluyen el documento de Monitoreo del Sistema, el cual contendrá los indicadores de uso del sistema y los cambios y mejoras propuestos.

3.3 Comparación y Selección de Metodología

Concluida la descripción de las metodologías más conocidas en proyectos de construcción de un DW, se muestra la comparación de las características más resaltantes en la Tabla 5.

Característica	Metodologías		
	Bill Inmon	**Ralph Kimball**	**Ramon Barquin**
Aspectos Metodología y Arquitectura			
Comité Directivo	Arriba-abajo (Top-Down)	Abajo-arriba (Bottom-Up)	Mixto según las necesidades
Estructura de la Arquitectura	DW para toda la empresa (atómica) Suministra datos a las base de datos departamentales	Data Mart para cada proceso del negocio; Consistencia a través de todos los Data Marts, mediante un data bus y dimensiones conformadas	DW para toda la empresa, a partir de Data Mart para cada proceso de negocio.
Complejidad del método	Compleja	Simple	Simple
Comparación con metodologías de desarrollo ya establecidas	Derivada de la metodología de Espiral	Proceso de 4 pasos: . Elegir el proceso de negocio. . Establecer el nivel de granularidad. . Elegir las dimensiones. . Identificar medidas y las tablas de hechos.	Considera el desarrollo en forma incremental, donde cada ciclo o iteración debe considerar: . Determinar Objetivos. . Análisis del riesgo. . Planificación. . Desarrollar y probar.
Modelamiento de la Data			
Discusión del diseño físico	Compleja	Simple	Simple
Orientación de la data	Orientada a la data	Orientada al proceso	Orientada a los requerimientos del negocio (procesos)
Herramientas	Tradicionales (ERD: Diagramas de Entidad Relación) (DIS: Data Item Sets)	Modelamiento dimensional, derivado del modelamiento relacional	Modelamiento dimensional utilizando herramientas de modelamiento tradicional
Acceso del usuario final	Baja	Alta	Alta
Filosofía			
Perfil de Usuario	Profesionales de TI	Usuarios finales	Usuarios finales
Objetivo	Dar una solución técnica eficiente basada en métodos y tecnologías de base de datos probados	Dar una solución fácil para el usuario final y pueda acceder a la data con una velocidad razonable	Brindar una solución dinámica que permita al usuario final acceder a la información en forma rápida
Otros aspectos			
Naturaleza de las necesidades de información de la organización	Estratégico	Táctico	Estratégico y Táctico
Requerimiento de integración de	Integración en toda la empresa	Áreas de negocios individuales	Áreas de negocios individuales

Característica	Metodologías		
	Bill Inmon	**Ralph Kimball**	**Ramon Barquin**
datos			
Estructura de Datos	Datos no medibles y datos que serán aplicados para satisfacer las múltiples y variadas necesidades de información.	Métricas de negocio, medidas de rendimiento y cuadros de mandos.	Datos medibles y datos que atenderán las múltiples necesidades de información.
Escalabilidad	Ámbito del crecimiento y la evolución de las necesidades son críticas	Necesidad de una alta volatilidad, se adapta a las necesidades dentro del ámbito limitado.	Fácil crecimiento según las necesidades de la organización.
Persistencia de los datos	Alta tasa de cambio en los sistemas de origen	Los sistemas de origen son relativamente estables	Las fuentes origen son relativamente estables
Dotación de personal y necesidades de capacidades	Grandes equipos de especialistas.	Pequeños equipos de generalistas.	Equipos medianos de especialistas y generalistas.
Tiempo de entrega	La Organización permite más tiempo para el arranque de la solución y disponibilidad de información.	La necesidad de la aplicación del DW es urgente.	La necesidad de la aplicación del DW es medianamente urgente.
El costo para implementar	Mayores costos al inicio y con menores costos posteriores de desarrollo de proyectos	Bajo costos de inicio, pero incrementando el costo de cada proyecto posterior, finalmente cuesta lo mismo.	Costo de inicio mediano, pero incrementando el costo de cada proyecto posterior, finalmente cuesta lo mismo.
Soporte Local	Escaso	Poco	Bueno
Entidad tiene personal con conocimiento de metodología	No disponible	No disponible	Disponible

Tabla 4: Metodologías para la construcción de un Data Warehouse

Fuente: Basado en Breslin, Mary (2004)

3.3.1 Calificación de Metodologías

Para la selección de la metodología se elaboró la Tabla 6 de calificación con las características analizadas anteriormente en función del proyecto a desarrollar y considerando para la calificación los valores 1, 2, 3 donde: 1: No conveniente, 2 Conveniente, 3 Muy Conveniente.

Característica	Metodologías		
	Bill Inmon	**Ralph Kimball**	**Ramon Barquin**
Enfoque general	2	2	3
Estructura de la Arquitectura	2	3	3
Complejidad del método	2	3	3
Comparación con metodologías de desarrollo ya establecidas	3	2	3
Discusión del diseño físico	2	3	3
Orientación de la data	2	2	3
Herramientas	3	2	2
Acceso del usuario final	2	3	3
Perfil de Usuario	2	3	3
Objetivo	3	3	3
Naturaleza de las necesidades de información de la organización	2	2	3
Requerimiento de integración de datos	3	2	2
Estructura de Datos	2	3	3
Escalabilidad	1	2	3
Persistencia de los datos	2	3	3
Dotación de personal y necesidades de capacidades	2	3	2
Tiempo de entrega	3	1	2
El costo para implementar	1	3	3
Soporte Local	1	2	3
Entidad tiene personal con conocimiento de metodología	1	1	3
Evaluación	**41**	**48**	**56**

Tabla 5: Calificación de las Metodologías

Fuente: Elaboración propia

3.3.2 Modelo a Aplicar

Del resultado de la calificación de las metodologías, las características de la información de los censos nacionales, su frecuencia de actualización, las condiciones para la explotación de la información y el dominio de la metodología en el entorno local, hacen que **se opte por la metodología de Ramón Barquin**.

Esta metodología es relativamente simple y su aplicación es más rápida cuando se trata de un primer proyecto de construcción DW, permite llegar a una solución completa en un menor tiempo minimizando los riesgos.

Capítulo IV: DESARROLLO DEL DATA WAREHOUSE

4.1 Planificación del Proyecto

4.1.1 Objetivos

Objetivo General

El "Proyecto Desarrollo Construcción del Data Warehouse de los Censos Nacionales del Instituto Nacional de Estadística e Informática de Perú", tiene como propósito la construcción de un repositorio centralizado de información censal, que facilite la explotación de datos para el análisis, de modo que todos los usuarios puedan acceder a la información censal a través de una aplicación en entorno Web.

Objetivo Específicos

- Generar reportes predefinidos totalmente dinámicos y si la información requerida no se encuentra dentro de dichos reportes, permitir al usuario diseñar nuevas consultas a la medida de sus necesidades, teniendo disponible para ello todos los atributos y medidas de análisis; además podrá incluir filtros y funcionalidades OLAP a sus reportes.

- Atender a una diversidad de público que no necesariamente cuenta con experiencia en el uso y análisis de información estadística, cuyos tipos y perfiles son: Tomadores de decisiones del gobierno central, tomadores de decisiones de

gobiernos regionales y locales, analistas de información de empresas privadas, académicos, profesionales, Estudiantes y público en general.

- Responder a los requerimientos de autoridades locales y regionales sobre los datos correspondientes a su jurisdicción así como a funcionarios del gobierno central atender demandas insatisfechas de la población.

4.1.2 Alcance del Proyecto

Análisis y diseño de un Data Warehouse de los Censos Nacionales utilizando la metodología de Ramón Barquín.

4.1.3 Descripción de actividades

Las actividades para el desarrollo del proyecto siguiendo la metodología de Barquin es la siguiente:

1) Desarrollo del Plan: Realizar la conceptualización, definición y alcance del proyecto, identificando las tareas asociadas al ciclo de vida, recursos, tiempos de duración, entre otros.

2) Definición de requerimientos de usuarios: Realizar consultas a los diferentes tipos de usuarios identificados y obtener sus requerimientos de información censal. Para esta actividad se realizaron talleres y reuniones de trabajo.

3) Diseño de la arquitectura del Data Warehouse. Desarrollar especificaciones técnicas para la infraestructura tecnológica del DW, teniendo en cuenta los requerimientos de información, la infraestructura instalada y las estrategias institucionales futuras. En Figura 12 se muestra la arquitectura del DW.

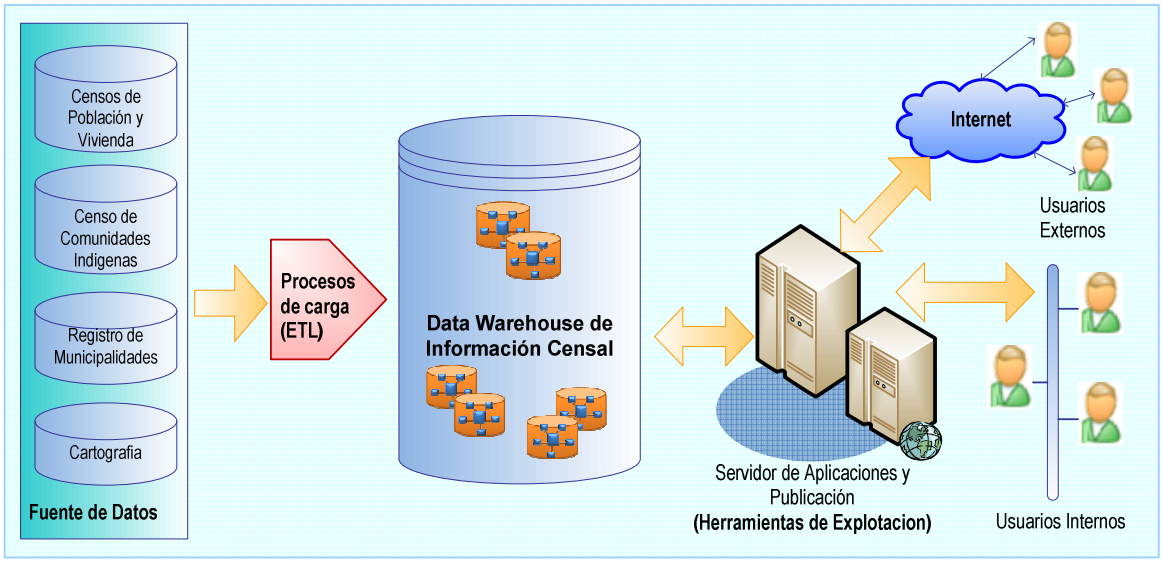

Figura 12: Arquitectura del Data Warehouse de los Censos Nacionales

Fuente: Elaboración Propia

La Arquitectura del DW de los Censos tiene los siguientes componentes:

- Fuentes de Datos, conformada por los Censos de Población y Vivienda 2007, Censo de Comunidades Indígenas, el Registro Nacional de Municipalidades y la Cartografía Censal, los cuales están en formato de archivos de texto, tablas en SQL Server y Oracle.

- Procesos de Carga, programas a desarrollar para la carga de datos utilizando la herramienta Keetle.

- Data Warehouse, repositorio centralizado que contiene el modelo multidimensional de datos, para lo cual utiliza Oracle como software de gestión de la base de datos.

- Herramientas de explotación, desarrollar una aplicación que proporcione funcionalidades avanzadas para el análisis de la información desde diversas dimensiones, usando la herramienta OLAP Microstrategy que permite crear informes y reportes de análisis a partir de los datos almacenados en el DW. Los usuarios pueden acceder a los datos por medio de un browser en Internet.

- Servidor de Aplicaciones, en este componente se instala y configura las aplicaciones que permitirán explotar la información del Data Warehouse y la publicación de la información por medio de Internet. Para lo cual se utiliza: JBoss, Microstrategy Web Server.

4) Instalación y pruebas del hardware y software para soportar el DW y las herramientas de acceso a la información.

5) Consistencia de datos, considerar la elaboración de indicadores para validar la exactitud de los datos.

6) Modelamiento dimensional, definir y diseñar las estructuras del modelo lógico y físico de las bases de datos.

7) Preparación y carga de datos, preparar la data para su carga en el modelo físico del repositorio de datos, las tareas se concentran principalmente en la extracción, formateo y carga de los datos.

8) Desarrollo de interfaz, aplicación de usuario final, en base a los requerimientos y prioridades establecidas. Considera el desarrollo de los esquemas de navegación y elaboración de los reportes pre establecido, así como las pruebas y correcciones necesarias.

9) Implementación del sistema, desplegar la tecnología, los datos y la aplicación de usuario final. Considerar capacitación, soporte técnico, estrategias de retroalimentación.

10) Monitoreo y crecimiento del sistema, continuar con los relevamientos de requerimientos en forma constante, en la revisión, modificaciones o cambios en los procedimientos de obtención de la información.

11) Gerencia técnica del proyecto, considerar la coordinación y dirección de las actividades técnicas del proyecto.

4.1.4 Organización Funcional del Proyecto

La organización funcional del proyecto se muestra en la Figura 13, donde se resaltan los equipos de trabajo tanto de INEI como de la empresa consultora.

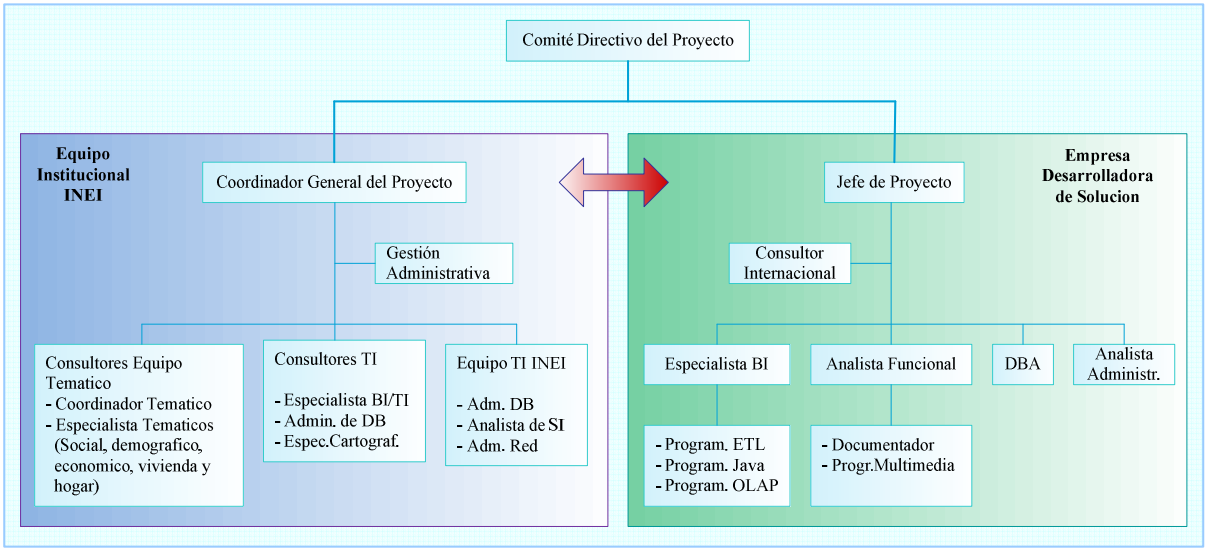

Figura 13: Organización del Proyecto Data Warehouse de los Censo Nacionales

Donde:

1) Conducción del Proyecto

Rol	Funciones
Comité Directivo	Conformado por funcionarios del INEI, identificados como principales usuarios de las direcciones que conducen y tienen la responsabilidad de producir y difundir la información censal. Lo preside el Jefe del INEI, quien da los lineamientos generales y la orientación a los objetivos del proyecto, también se encarga de asegurar los recursos necesarios para la ejecución del mismo.

Tabla 6: Rol de Comité Directivo

2) Equipo Institucional INEI

Rol	Funciones
Coordinador General	• Encargado de Planificar, controlar, coordinar, articular e integrar el desarrollo de las actividades y productos. • Promover la coordinación e integración de los equipos del INEI y de

Rol	Funciones
	la empresa consultora para lograr que el proyecto se desarrolle con los más altos estándares de eficiencia, eficacia y calidad.
Asistente Administrativo	Responsable de la gestión administrativa del proyecto, realiza el monitoreo y seguimiento administrativo de todas las actividades del proyecto DW.
Consultores Equipo Temático	
Coordinador Temático	• Dirige, coordina y supervisa las actividades de los especialistas temáticos en las diferentes disciplinas y asegurar el logro de los objetivos del proyecto. • Conjuntamente con los especialistas temáticos definir los cuadros e indicadores estadísticos, así como los criterios para aplicar el Secreto Estadístico. • Aprobar la valides de la calidad de la información cargada en el DW.
Especialista en temas Sociales, Económicos, Demográficos, Vivienda y Hogar	• Responsables de asegurar que la información a ser publicada responda a los requerimientos de los usuarios finales con la calidad requerida. Asegurando la aplicación de metodologías estadísticas para la generación de cuadros y el cálculo de indicadores.
Consultores TI	
Especialista BI/TI	• Responsable de definir los aspectos metodológicos del proyecto y supervisar del desarrollo de las actividades. • Elaborar protocolos para la evaluación de avances del proyecto y su aceptación.
Administrador de BD	• Responsable de aprobar el diseño e implementación del modelo multidimensional del DW. • Supervisar la instalación, configuración y afinamiento de la base de datos.
Especialista Cartográfico	• Responsable de definir aspectos geográficos y cartográficos a ser incluidos en el DW.
Equipo TI INEI	
Administrador de BD	• Responsable de trasladar el conocimiento de las bases de datos institucionales relacionadas al diseño del DW. • Responsable del mantenimiento y crecimiento futuro del DW. • En coordinación con el Consultor Administrador de BD, validar la arquitectura, el diseño del modelo de datos del DW.

Rol	Funciones
Analista de Sistemas	• Responsable de trasladar el conocimiento de las reglas de negocios para su implementación en los procesos de generación de reportes. • Apoyar al Consultor de BI/TI en detallar y validar las especificaciones técnicas para aplicar las reglas del negocio. • Responsable de la sostenibilidad, crecimiento y administración de la solución.
Administrador de Red	• Responsable del crecimiento, sostenibilidad y administración de la plataforma tecnológica del INEI que soporta el DW.

Tabla 7: Rol del Equipo Institucional INEI

3) Empresa Consultora

Rol	Funciones
Jefe de Proyecto	• Planificar, organizar, dirigir, coordinar, controlar y administrar el proyecto. • Seleccionar y proveer los recursos humanos y tecnológicos necesarios para el desarrollo de las actividades. • Coordinar, proporcionar y recoger informaciones e inquietudes de los usuarios y organismos participantes en el proyecto.
Consultor Internacional	• Validar el plan del proyecto y orienta la aplicación de la metodología de construcción del DW.
Especialista BI	• Encargado del liderazgo tecnológico del proyecto y apoya al Jefe de Proyecto en la definición de los objetivos, alcances, beneficios, plan de acción sugeridos y perfil del equipo. • Participar en el análisis de requerimientos, el desarrollo del cronograma preliminar, la asignación de recursos y conduce el control de calidad de la solución. • Definir los lineamientos y estrategias para el desarrollo del proyecto. Propone el diseño de la arquitectura de la solución, buscando el adecuado uso de la tecnología.
Programador ETL	• Experto en procesamiento de grandes volúmenes de datos • Encargado de desarrollar los programas para la carga de datos y documentar los procesos a ejecutar por operadores
Programador	• Especialistas en la plataforma de desarrollo y/o plataforma de BI

Rol	Funciones
Java	seleccionada. • Encargado de las labores de análisis, diseño, construcción, e implementación de las herramientas para acceder al DW.
Programador OLAP	• Responsable de diseñar la aplicación final con característica de OLAP para acceso a consultas en línea.
Analista Funcional	• Responsable de elaborar los estándares y definir las metodologías de desarrollo de sistemas a utilizarse. • Realiza el análisis y diseño del aplicativo de acceso a la información. • Realizar las pruebas unitarias de los programas y apoyar en la integración e implantación del sistema.

Tabla 8: Rol de Empresa Consultora

4.1.5 Cronograma de Actividades

En la Figura 14 se presenta el cronograma de alto nivel, que considera el desarrollo del DW de los Censos Nacionales que inicialmente incluye el Censos de Población y Vivienda y el Censo de Comunidades Indígenas materia de esta tesina; el crecimiento que considera la incorporación del Censo Económico 1994 y del 2008, el Censo Universitario, el Censo Agropecuario 1994 y del 2012. Cabe precisar que el proceso de crecimiento de este DW debe ser una actividad continua.

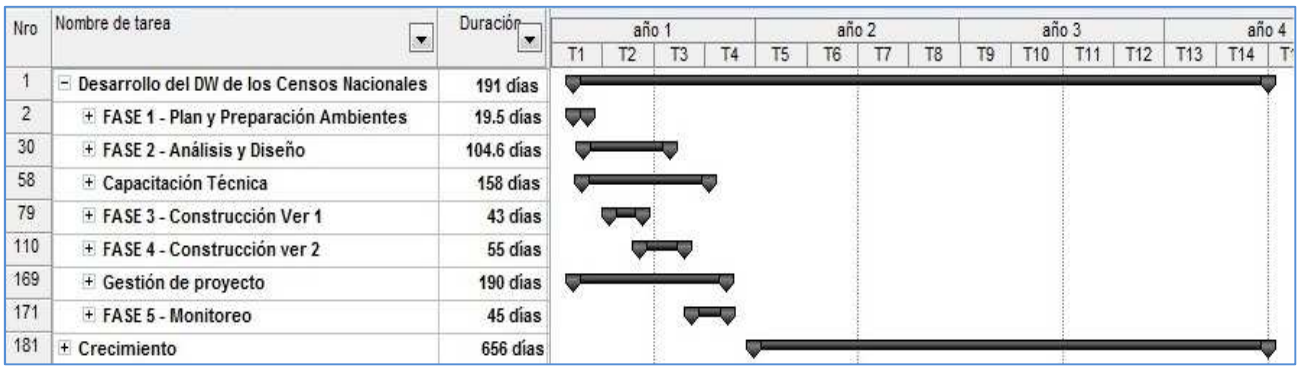

Figura 14: Cronograma a alto nivel

4.1.6 Inversión

En la Tabla 10 se presenta la inversión total del proyecto para la construcción del DW del Censo Nacional de Población y Vivienda.

Nº	ACTIVIDADES	TOTAL $
1	Capacitaciones, talleres y reuniones de trabajo con usuarios potenciales del DW	93,881
2	Definición de requerimientos de usuarios (Consultores temáticos)	194,064
3	Consistencia y preparación de información censal	198,469
4	Gestión del Proyecto	51,974
5	Construcción de DW (Consultores TIC)	414,578
6	Software (licencias por 3 años)	274,000
7	Hardware	173,568
	TOTAL $	**1,400,534**

Tabla 9: Inversión realizada en el Proyecto

Fuente: INEI (2008)

4.1.7 Límites de la Solución

La solución integra y administra la información de los Censos de Población y Vivienda 2007, el censo de Comunidades Indígenas e información relacionada a ambas operaciones. Dicha información tiene una nueva versión cada 10 años.

La información a la que accederá el usuario corresponde a los datos básicos y a los niveles geográficos determinados por la política institucional de difusión de datos y resguardo de la confidencialidad de los mismos, los cuales estarán plasmados en el sistema.

El usuario final accede al sistema vía Internet, por lo tanto está limitado a los lugares que cuentan con este servicio.

La función de modificación y actualización de los modelos de datos, la agregación de nuevos modelos y el acceso a las consolas de administración del sistema son de acceso restringido al personal especializado.

El acceso a la información censal no pública está restringido por la política de difusión institucional y el aseguramiento del resguardo del secreto estadístico.

4.1.8 Factores Críticos de Éxito

Durante la fase de anteproyecto se identificaron las funciones, actividades y aspectos en los que no se puede fallar, porque de hacerlo, se compromete la viabilidad del proyecto. A continuación se detallan los factores críticos de éxito identificados:

- **Compromiso de la alta dirección de INEI**
  Es de mucha importancia la participación de la alta dirección del INEI para lograr un compromiso adecuado por parte de los colaboradores del INEI, así como para el apoyo en la toma de decisiones sobre temas de alto impacto en el proyecto.

- **Cumplimiento de las fechas límite en lo referente a contrataciones y licitaciones**
  Una de las principales causas generadoras de retrasos en el proyecto son los procesos de contrataciones y licitaciones. Se deberán de tomar las medidas correspondientes para agilizar en todo lo que se puedan estos procesos, para que no exista la necesidad de mover fechas en los cronogramas de trabajo.

- **Disponibilidad por parte de los recursos humanos de INEI**
  La participación del personal del INEI, tanto técnico como funcional. Es necesario identificar la carga real de trabajo de estas personas para conocer el tiempo que pueden dedicarle al proyecto.

- **Autorización para tomar decisiones de los representantes de las diferentes Direcciones Nacionales y Direcciones Técnicas del INEI**
  Es fundamental que el personal asignado por las diferentes Direcciones, puedan tomar decisiones funcionales y técnicas a fin de optimizar el tiempo firmando los acuerdos.

- **Calidad de datos del Censo de Población y Vivienda 2007**

  Este es el factor crítico más importante. El nivel de calidad de información que se le entregue al usuario final deberá ser exacto. Si esto falla, el sistema, bueno o malo, deja de tener sentido.

- **Interfaz amigable**

  El principal objetivo del desarrollo de la interfaz es que sea muy fácil de utilizar e intuitiva. Para esto, durante la etapa de diseño se deberán de validar desde diferentes puntos de vista todos los prototipos elaborados, a fin de identificar el mejor desde el punto de vista del usuario.

- **Generación de sinergias entre los miembros del equipo del proyecto**

  Un adecuado trabajo en equipo incrementa enormemente las probabilidades de éxito de todo proyecto. Para esto se deben establecer los objetivos comunes de todo el equipo (INEI, consultores externos, proveedores) y definir los roles, funciones y responsabilidades de cada uno. Siempre se debe buscar un adecuado ambiente de trabajo donde se pueda generar el compromiso y confianza por parte de todos los miembros del equipo.

- **Comunicación**

  El equipo del proyecto está conformado por un número significativo de profesionales, con diferentes perfiles, especialidades, edades, objetivos, modelos mentales, etc. La comunicación se vuelve un factor crítico de éxito debido a que por lo heterogéneo del grupo humano, los mensajes no podrían estar llegando de la mejor manera posible. Debemos de crear los canales de comunicación adecuados para que tanto los mensajes, como la información, fluyan de la manera más adecuada posible.

- **Uso de metodologías, estándares y adecuada gestión de proyectos**

  El uso de una metodología para el proyecto ayuda a llevar un mejor control del mismo y mantiene alineados hacia cumplimiento de los objetivos del

proyecto. Esto sumado a una eficaz gestión de proyectos, permite disminuir las probabilidades de fracaso del proyecto.

4.2 Requerimiento de Usuarios

4.2.1 Usuarios identificados

El DW busca atender a una diversidad de público que no necesariamente cuenta con experiencia en el uso y análisis de información estadística. Los principales tipos de usuarios se muestran en la Figura 15.

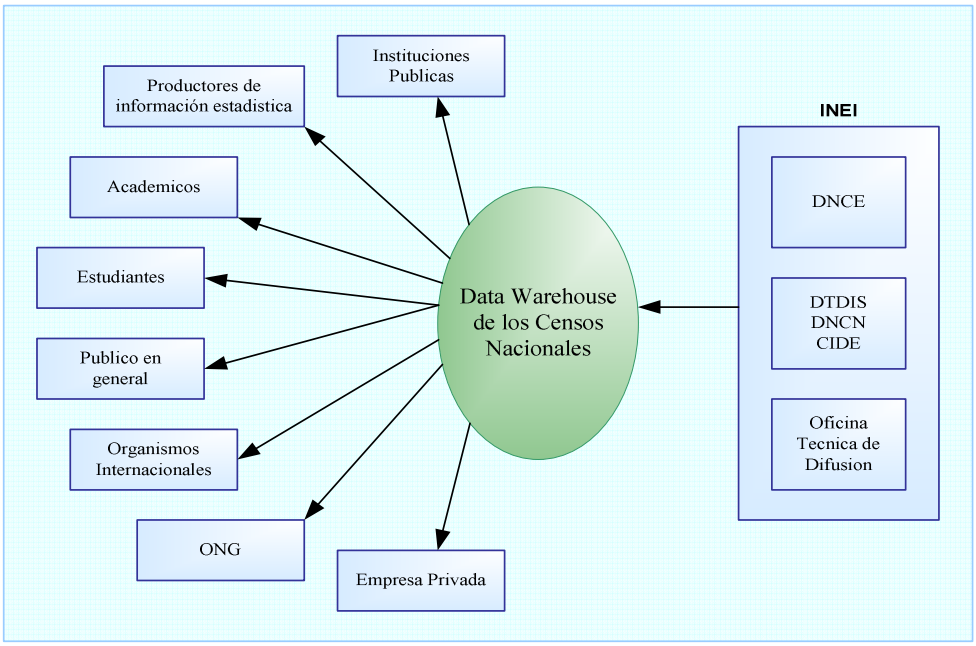

Figura 15: Tipos de usuarios del Data Warehouse de los Censos Nacionales

Especial atención merecen las necesidades de usuarios calificados, en el sentido de las particulares exigencias que se darían por la especialización que tienen en el manejo de información estadística, como son los usuarios internos del INEI, productores y usuarios de información estadística del sector público y analistas de información, investigadores y académicos.

Para ello se convocó a los funcionarios claves de diversas entidades públicas y privadas, a quienes se les brindó capacitaciones sobre las características y beneficios que provee el DW para mejorar los procesos de toma de decisiones; también se realizaron talleres

para el manejo y utilización de información censal y simultáneamente se recopiló los requerimientos de información de estas entidades. En la Tabla 11 se muestra las instituciones que participaron en el proceso de identificación de requerimientos.

Nº	NOMBRE DE ENTIDAD	Nº Participantes	%
1	Agencia Peruana de Cooperación Internacional	1	1
2	Asamblea Nacional de Rectores (ANR)	1	1
3	Consejo Nacional de Ciencia y Tecnología (CONCYTEC)	1	1
4	Consejo Nacional del Ambiente	2	3
5	Fondo de Cooperación para el Desarrollo Social (FONCODES)	1	1
6	Fondo Nacional de Desarrollo Pesquero (FONDEPES)	1	1
7	Instituto del Mar Peruano (IMARPE)	2	3
8	Instituto Nacional de Cultura (INC)	1	1
9	Instituto Nacional de Recursos Naturales (INRENA)	4	6
10	Instituto Nacional de Defensa Civil	1	1
11	Instituto Peruano del Deporte (IPD)	1	1
12	Ministerio de Agricultura	2	3
13	Ministerio de Comercio Exterior y Turismo	3	4
14	Ministerio de Defensa	1	1
15	Ministerio de Economía y Finanzas	6	9
16	Ministerio de Educación	1	1
17	Ministerio de Energía y Minas	2	3
18	Ministerio de Justicia	1	1
19	Ministerio de la Mujer	3	4
20	Ministerio de la Producción	5	7
21	Ministerio de Relaciones Exteriores	1	1
22	Ministerio de Trabajo y Promoción del Empleo	5	7
23	Ministerio de Transportes y Telecomunicaciones	2	3
24	Ministerio de Vivienda Construcción y Saneamiento	6	9
25	Ministerio del Interior	1	1
26	Ministerio Público	2	3
27	Presidencia del Consejo de Ministros	3	4
28	Prog.Nac. de Manejo de Cuencas Hidrográficas y Conservación de Suelos (PRONAMACHCS)	2	2
29	Registro Nacional de Identificación y Estado Civil (RENIEC)	1	1
30	Superintendencia de Banca, Seguros y AFP	1	1
31	Seguro Integral de Salud (SIS)	1	1
32	Superintendencia Nacional de los Registros Públicos (SUNARP)	1	1
33	Super Intendencia Nacional de Administración Tributaria (SUNAT)	1	1
	Total	**67**	**100**

Tabla 10: Usuarios participantes por Entidad

Fuente: INEI, 2008

4.2.2 Definición de Requerimientos

Los principales requerimientos identificados son:

1) Mejorar el tiempo de respuesta al usuario cuyos requerimientos toman cerca de dos semanas o incluso más.
2) Eliminar posibles diferencias en los resultados mostrados en los diversos aplicativos y publicaciones.
3) Facilidades para la interpretación de la información apoyada de metadatos.
4) Proporcionar cuadros estadísticos dinámicos con funcionalidades para realizar análisis, que permitan ver la información a nivel macro e ir ingresando a niveles más detallados, considerando:

- **Reportes a todo nivel de detalle de la Ubicación Geográfica**

  De la Figura 16 se observa que más del 80% de usuarios requiere que los diversos cuadros estadísticos sean generados a nivel de departamento, provincia y distrito lo que permite conocer a nivel macro la situación del país. Para estudios más especializados requieren que los cuadros sean desagregados a nivel centros poblados, áreas rurales y urbanos y de manzanas.

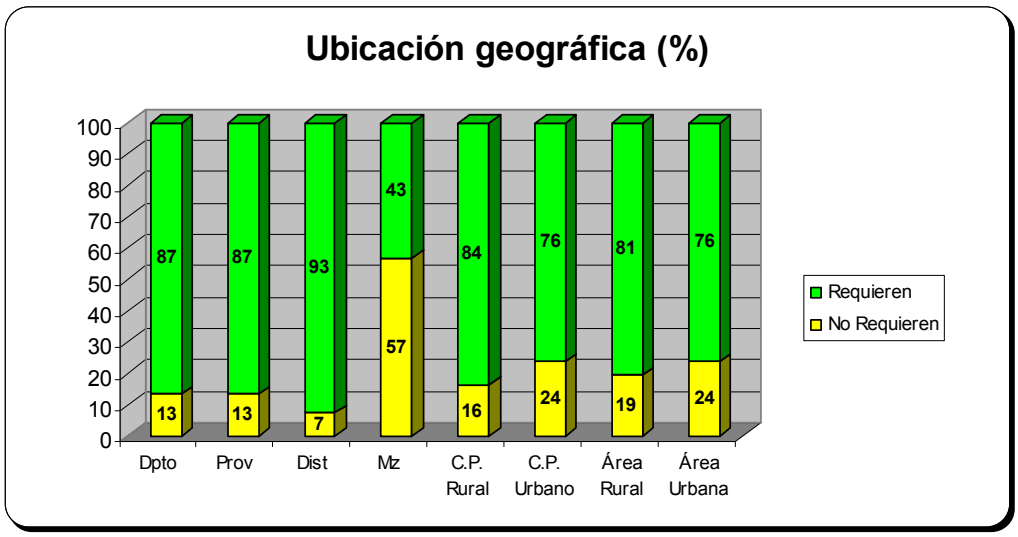

Figura 16. Porcentaje de usuarios que requieren información por nivel Geográfico

Fuente: INEI (2008)

- **Reportes sobre características de la vivienda**

  La información sobre las características de las vivienda es de gran importancia para cerca del 60% de usuarios claves de las instituciones públicas, quienes

analizan la *condición de ocupación de las viviendas*, *material de construcción y servicios básicos*, todo ello a fin de priorizar las instalación de nuevos servicios donde no tienen o impulsar programas de apoyo para construcción de viviendas, entre otros. En la Figura 17 se presenta los requerimientos relacionadas con las viviendas.

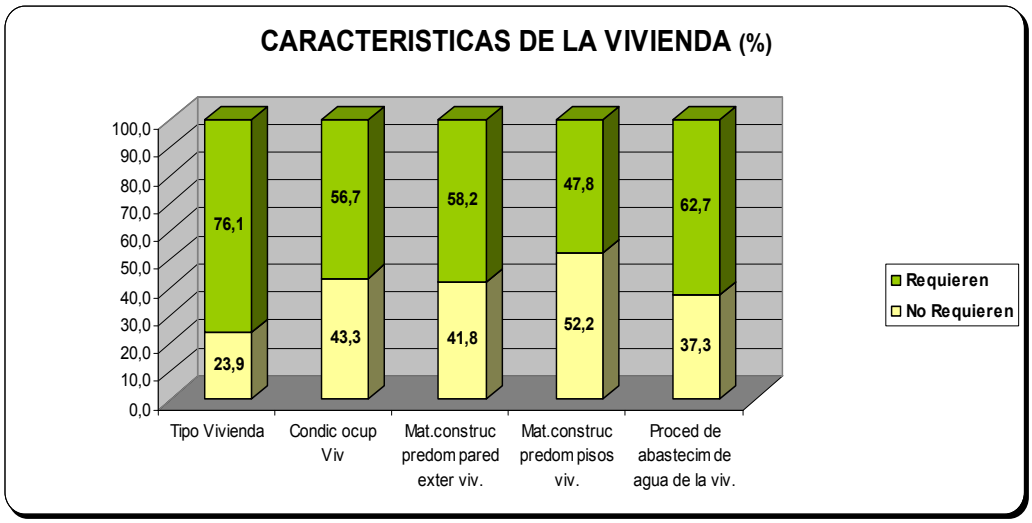

Figura 17: Porcentaje de usuarios que requieren información sobre vivienda

Fuente: INEI (2008)

- **Reportes sobre características demográficas**

  En la Figura 18, que más del 80% en promedio de usuarios consultados, manifiesta la necesidad de contar con información detallada sobre las características demográficas de la población que permite planificar y ejecutar programas y políticas de desarrollo a nivel nacional, departamental, provincial y distrital. La información de población por sexo y edad permite estimar la demanda de los servicios básicos de salud, educación y empleo.

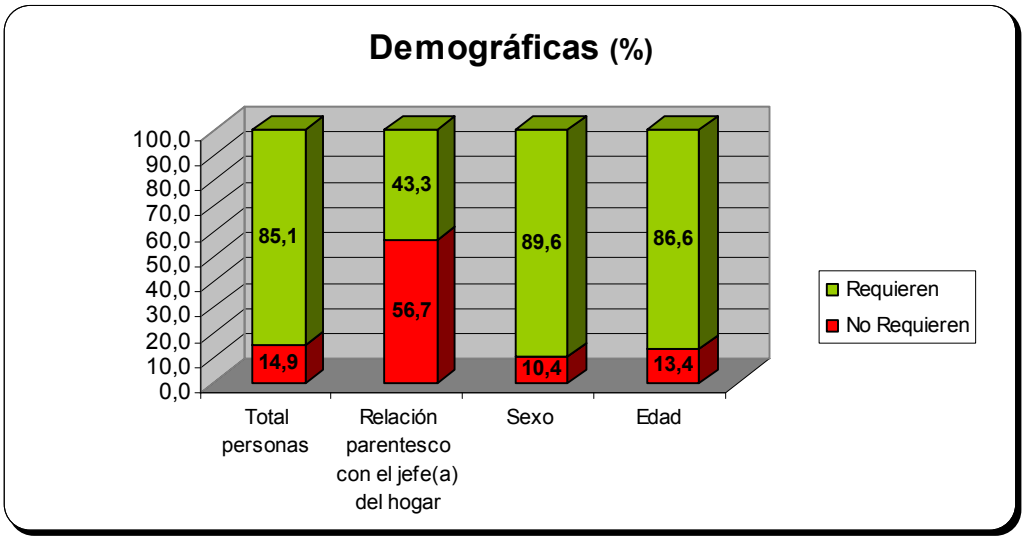

Figura 18: Usuarios que requieren información socio demográfica

Fuente: INEI (2008)

- **Reportes sobre características demográficas**

  De la Figura 19 se deduce que más del 60% de usuarios consultados, plantea la necesidad de contar con información detallada sobre cultura y educación, con lo cual se puede calcular la tasa de analfabetismo y evaluar el capital educativo de un país. Así mismo, permite evaluar, diseñar e implementar políticas y programas de alfabetización.

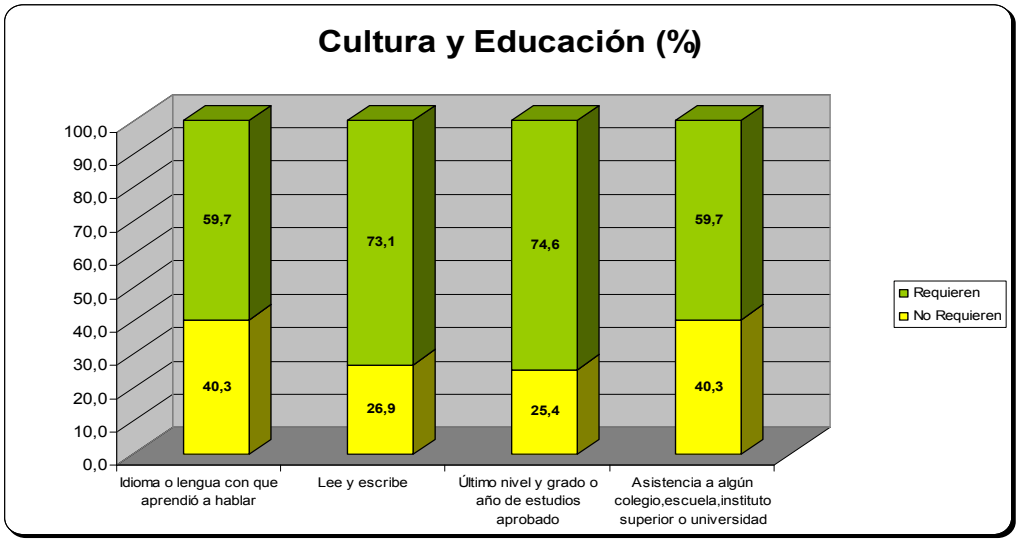

Figura 19: Usuarios que requieren información sobre cultura y educación

Fuente: INEI (2008)

- Reportes sobre características Económico Laboral

En Figura 20, se observa que el 88% de usuarios solicita conocer la población económicamente activa (PEA); el 78% plantea la necesidad de conocer la ocupación principal que realiza. Todo ello a fin de evaluar la participación de la fuerza laboral, Requerimientos de empleo y conocer la estructura económica del país por sector económico.

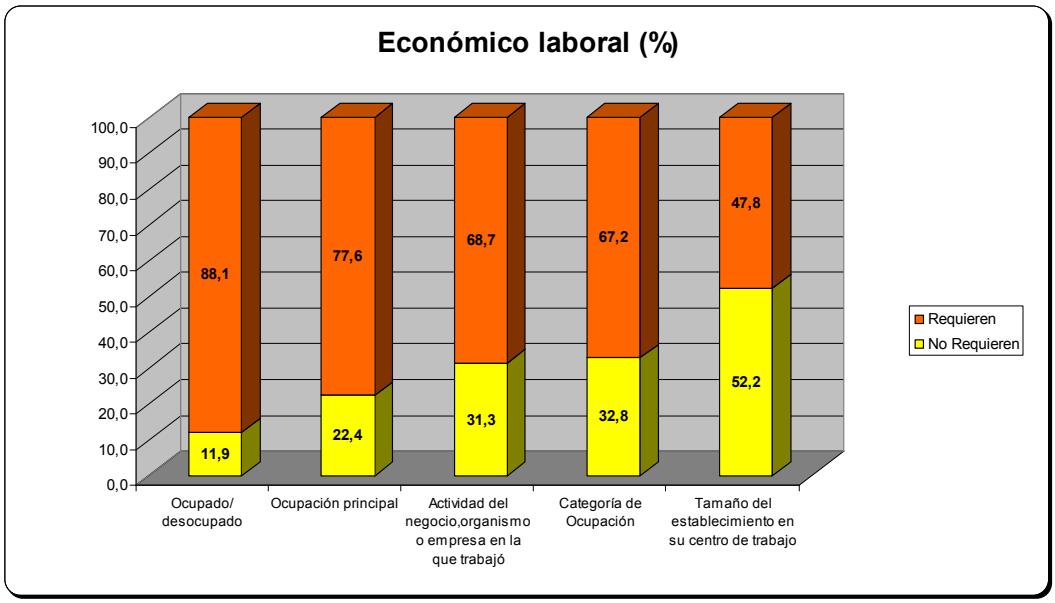

Figura 20: Usuarios que requieren información económica

Fuente: INEI (2008)

4.3 Modelamiento Dimensional

4.3.1 Modelo conceptual de datos

- **Modelo General**

  En base a los cuestionarios censales utilizados para el levantamiento de información, se presenta en Figura 21 el modelo del DW a nivel macro en la cual se identifican las entidades que lo conforman.

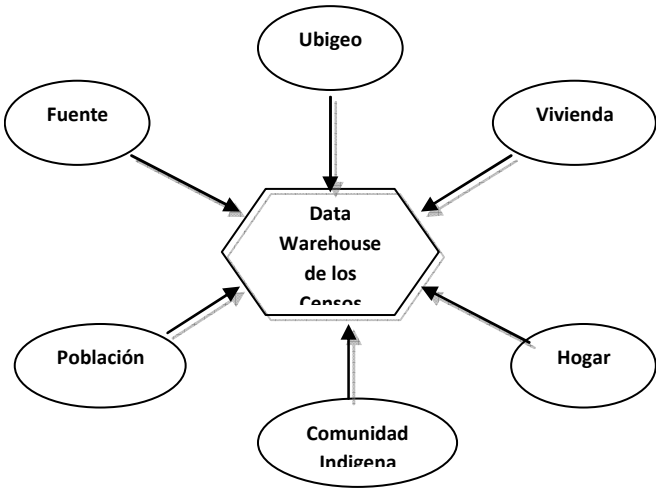

Figura 21: Modelo general del DW de los Censos

En Tabla 12 se muestra la descripción de cada una de las entidades

Entidad	Descripción	Ejemplo
Fuente (Censo)	Es el conjunto de actividades destinadas a obtener información de las principales características de todos los habitantes y de las viviendas que se encuentran en el territorio nacional.	CPV 2007
Ubigeo	Ubigeo son las siglas oficiales para Código de Ubicación Geográfica, que usa el INEI para codificar las circunscripciones territoriales del Perú.	15 Departamento de Lima
Vivienda	Es una edificación o unidad de edificación independiente construida, adaptada o convertida para ser habitada por una o más personas en forma permanente o temporal.	Vivienda Particular, Personas presentes, con alumbrado, etc.
Hogar	Es la persona o conjunto de personas, sean parientes o no, que ocupan en su totalidad o en parte una vivienda; comparten al menos las comidas principales y atienden en común otras necesidades básicas, con cargo a un presupuesto común.	Con TV, PC, Energía, Discapacidad, etc.
Población	Es la persona que vive o se encuentra temporalmente en la vivienda, del cual se requiere obtener los datos respectivos.	Parentesco, sexo, idioma, religión, DNI, educación, etc.
Comunidad Indígena	Son grupos tribales de la Selva y Ceja de Selva, constituidos por un conjunto de familias. Están	TAMPE, Pueblo AGUARUNA

Entidad	Descripción	Ejemplo
	relacionado al CPV 2007	

Tabla 11: Entidades del modelo de datos

- **Análisis de Viviendas**

  En la Figura 22 se muestra el modelo se basa en los datos del censo exclusivo para viviendas. Por otra parte se detallan las entidades que forman parte del análisis.

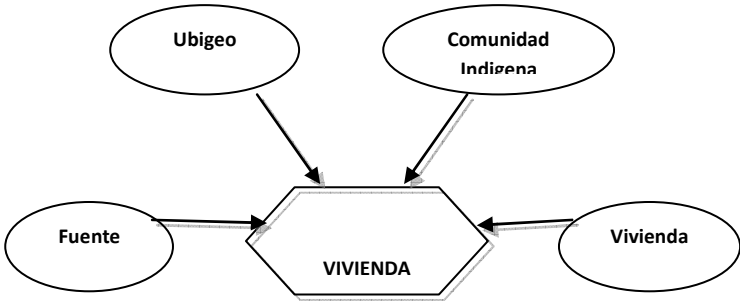

Figura 22: Modelo para la entidad vivienda

- **Análisis de Hogares**

  En la Figura 23 se muestra el modelo de la entidad Hogar. Sirve para el análisis exclusivo de hogares dentro de una vivienda. Se muestran también las entidades que forman parte del análisis.

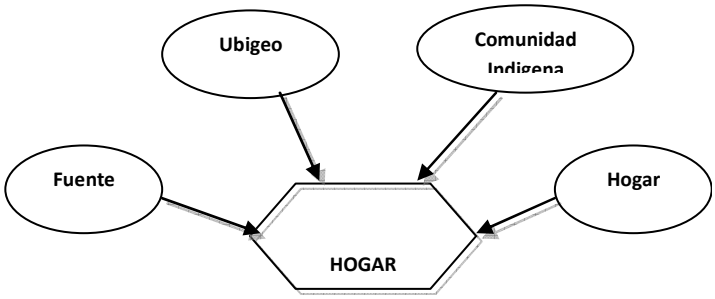

Figura 23: Modelo para la entidad hogar

- **Análisis de Población**

  En la Figura 24 se muestra el modelo para la enditad población con las otras entidades que forman parte del análisis.

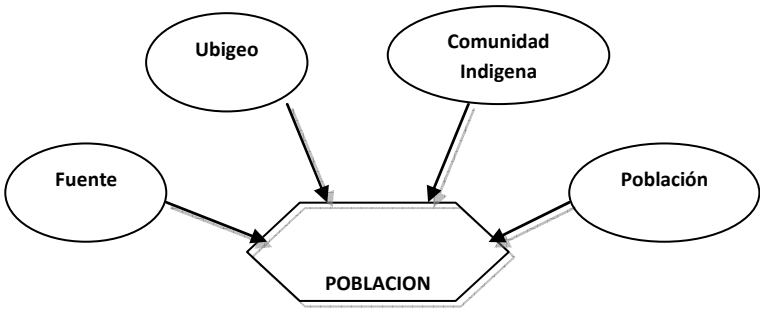

Figura 24: Modelo para la entidad población

- **Análisis de Comunidades Indígenas**

  Esta sección comprende el análisis de las comunidades indígenas por ubicación geográfica y con las características que las identifica como se muestra en la Figura 25. Asimismo, se detallan las entidades consideradas para este modelo.

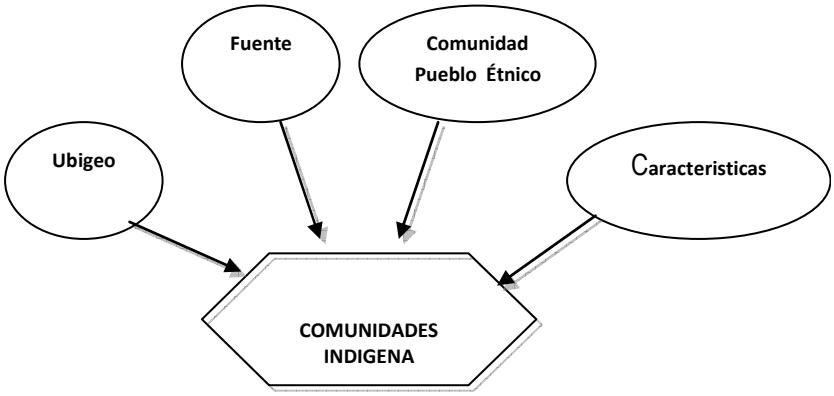

Figura 25. Modelo para la entidad Comunidad Indígena

4.3.2 Diagramas Funcionales

El alcance de los requerimientos de información planteados ha sido generado a partir de las reuniones de trabajo sostenidas con usuarios internos y externos del INEI, de quienes se ha podido recopilar y definir todas las dimensiones y sus métricas que forman parte del análisis de información.

A continuación se mostraran los diagramas funcionales (diagramas Star Net), que permite identificar las todas las métricas y dimensiones que se utilizarán en el análisis de información. A partir de los cuales se podrán construir los diferentes reportes y gráficos de apoyo a la gestión.

- **Análisis del Modelo General**

  Este modelo se basa en la información del censo para el análisis combinado de Viviendas, Hogares y Población. Es decir, este enfoque permite determinar que hay una serie de análisis de información que se pueden desprender, en base a las tres grandes variables de la solución de acuerdo a lo mostrado en la Figura 26.

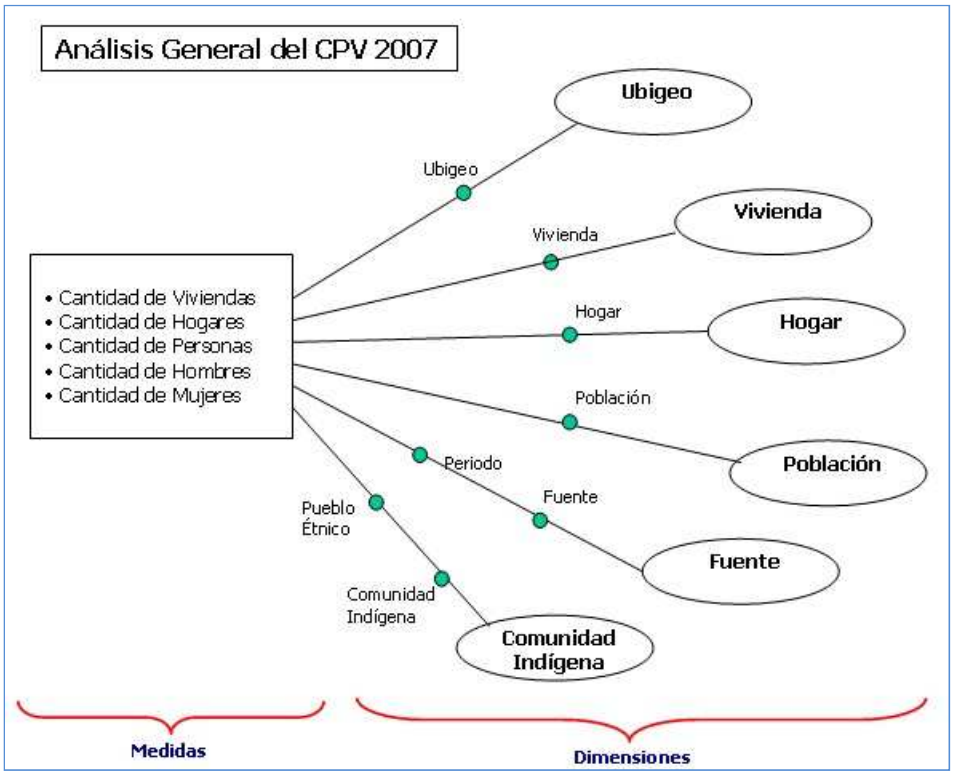

Figura 26: Análisis del modelo General del DW

A continuación presentaremos en forma independiente los diagramas correspondientes a Viviendas, Hogares, Población e Indicadores, cada uno con las medidas y dimensiones correspondientes.

- **Análisis de Vivienda**

  Este modelo permitirá obtener información estadística de las cantidades de viviendas, teniendo la posibilidad de extraer una serie de decisiones a partir de variables como Tipo de Vivienda, Condición de Vivienda, Material de Construcción, etc.

A continuación se muestra el modelo y se detallan cada uno de los ítems que conforman las dimensiones y medidas tal como se muestra en la Figura 27.

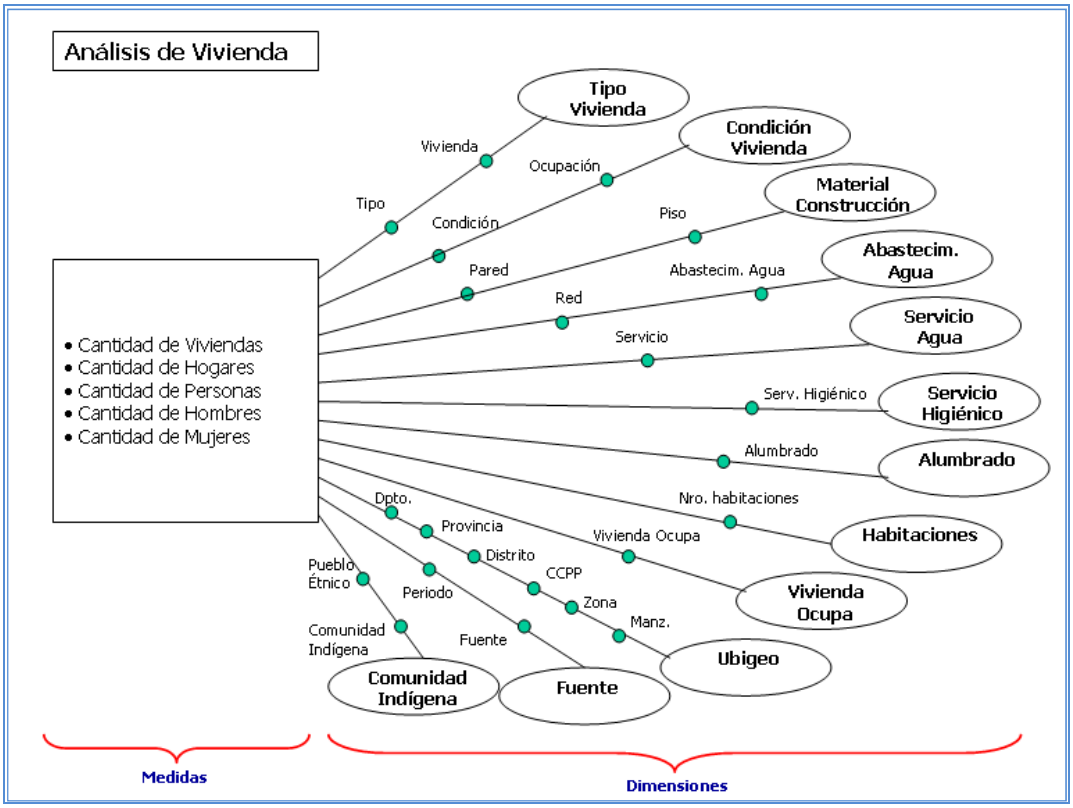

Figura 27: Modelo a detalle de la entidad vivienda

- **Análisis de Hogar**

  Este modelo permitirá obtener información estadística de las viviendas, teniendo la posibilidad de analizar las combinaciones de variables por cada medida que se tiene. Por ejemplo cruzar equipos, energía, cantidad de hogares, ubigeo y fuente.

  En la figura se muestra el modelo y se detallan cada uno de los ítems que conforman las dimensiones y medidas de acuerdo a lo mostrado en la Figura 28.

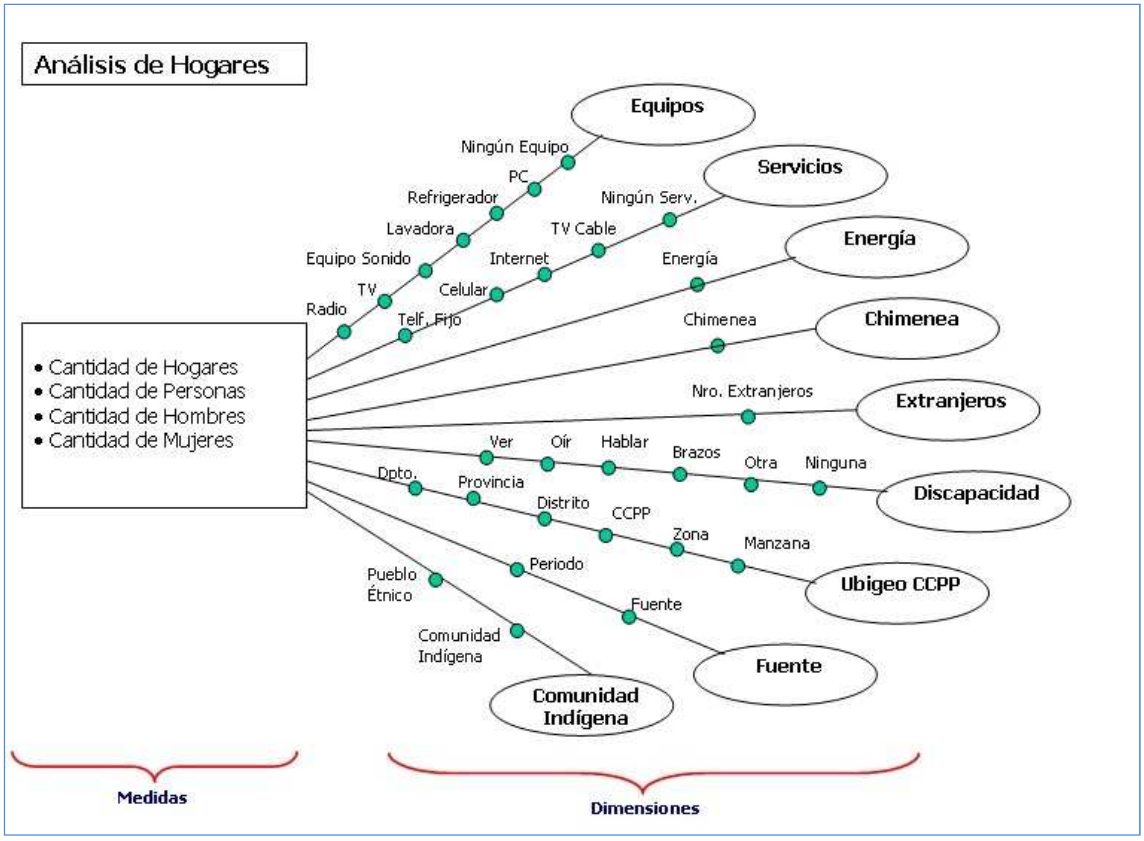

Figura 28: Modelo a detalle de la entidad hogar

- **Análisis de Población**

  Este modelo permitirá obtener información estadística de la población, teniendo la posibilidad de analizar las combinaciones de variables por cada medida que se tiene. Por ejemplo cruzar sexo, edad, trabajo, cantidad de persona, ubigeo y fuente.

  En la Figura 29 se muestra el modelo y se detallan cada uno de los ítems que conforman las variables y medidas.

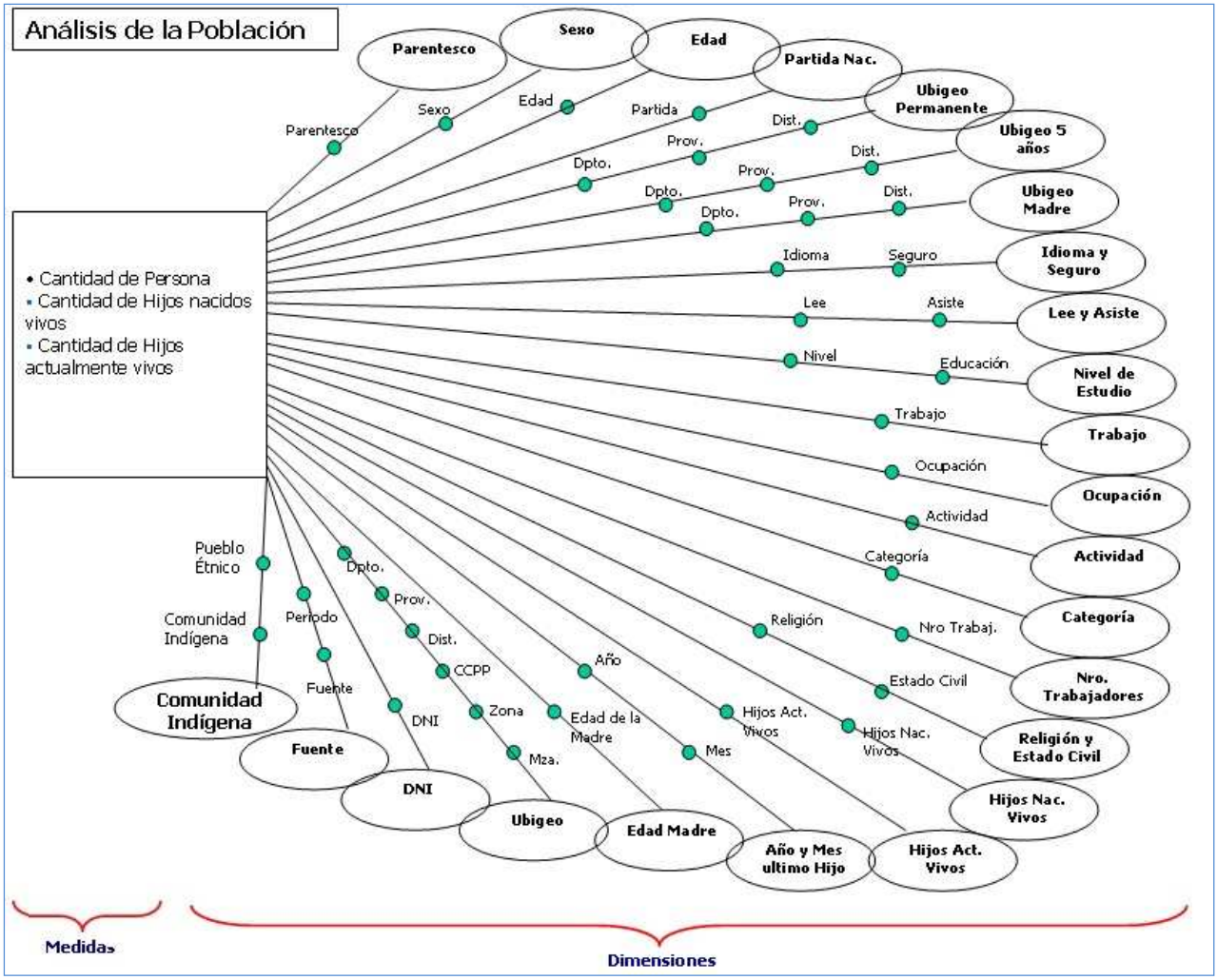

Figura 29. Modelo a detalle de la entidad Población

- **Análisis de Indicadores**

  Este modelo está orientado al análisis de indicadores proporcionados por el INEI, cuyo análisis será por el Ubigeo, por la Clase, por Are temática (Empleo, Salud, etc.) o por sub área temática, para luego llegar al indicador en sí. En dicho diagrama se podrá apreciar cada una de las variables del modelo así como la medida que contiene el valor del indicador. Por ejemplo permite cruzar indicador, ubigeo, clase, fuente y valor del indicador.

  En Figura 30 se muestra el modelo y se detallan cada uno de los ítems que conforman las variables y medidas.

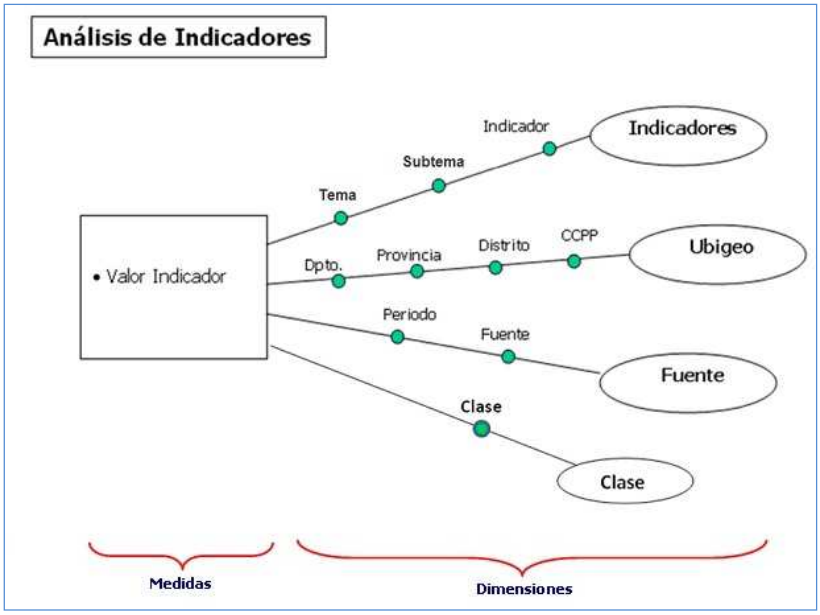

Figura 30. Modelo a detalle de los Indicadores

4.3.3 Modelo Físico

En la Figura 31 se muestra la relación de dimensiones a nivel atómico o de detalle, cabe resaltar que este modelo sirve como base para la generación de las tablas agregadas.

También se aprecia las 3 tablas de hechos: F_VIVIENDA, F_HOGAR y F_POBLACION y el resto son las tablas de dimensiones.

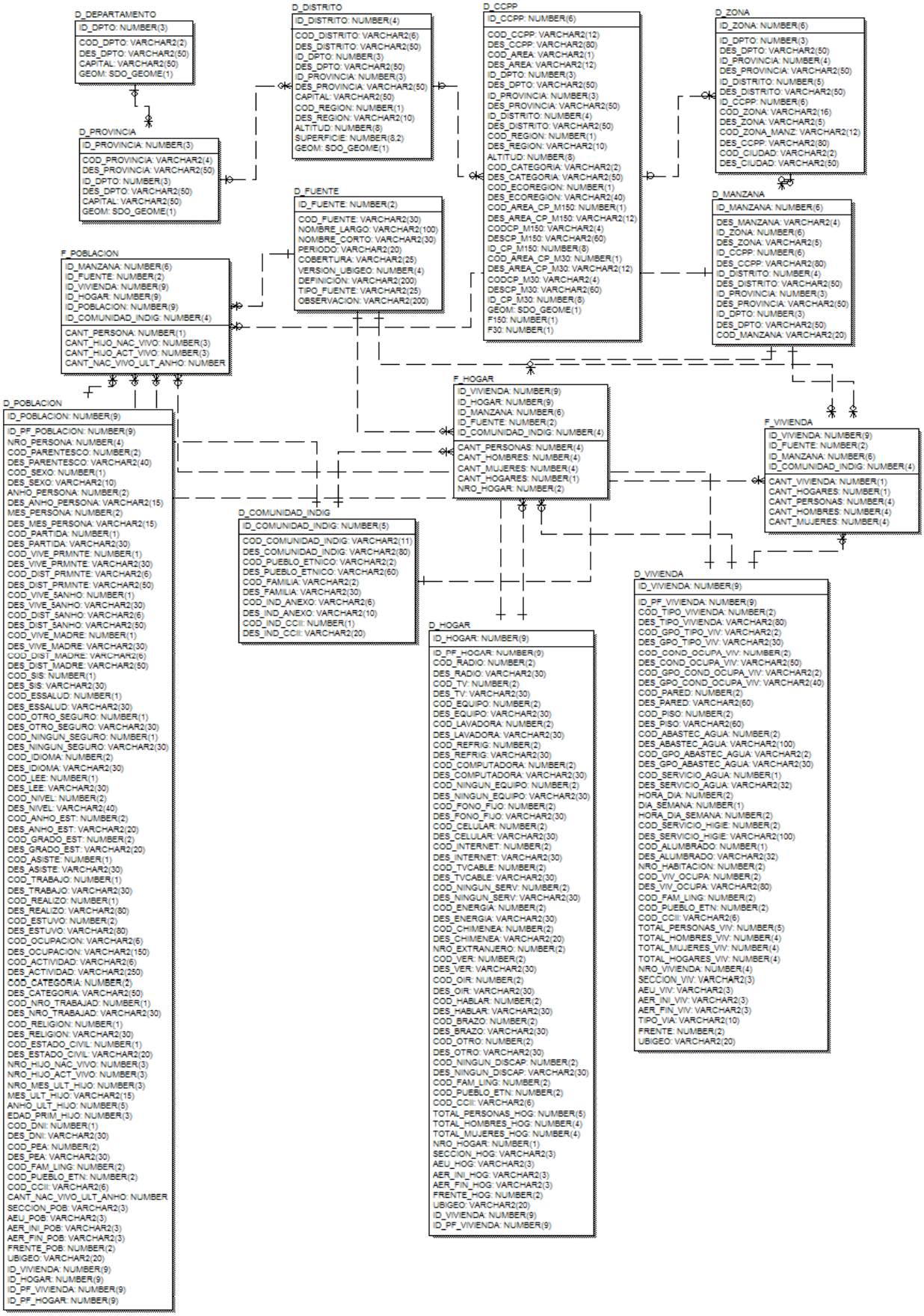

Figura 31: Modelo de datos físico

- **Dimensiones**

  En Tabla 13 se muestra la relación de dimensiones identificadas durante el análisis de la información censal.

Dimensión	Descripción
Departamento	Contiene atributos referentes al departamento, que es la subdivisión geográfica del Perú, que está conformado por 25 departamentos incluyendo Callao.
Provincia	Contiene atributos referentes a la Provincia, que la subdivisión geográfica de los departamentos, el Perú cuenta en la actualidad con 195 provincias.
Distrito	Contiene atributos referentes al Distrito, que es la subdivisión geográfica de las provincias y es la menor división político administrativo del país y comprende el conjunto de centros poblados urbanos y rurales. El Perú cuenta en la actualidad con 1834 distritos.
Centro Poblado	Contiene atributos referentes al Centro Poblado, que es la subdivisión geográfica de los distritos. El Perú cuenta en la actualidad con 98011 centros poblados entre urbanos y rurales.
Zona	Contiene atributos referentes a la zona censal, que es la subdivisión de los centros poblados urbanos. El Perú cuenta en la actualidad con 6792 Zonas.
Manzana	Contiene atributos referentes a la Manzana censal, que es subdivisión de las zonas. El Perú cuenta en la actualidad con 312078 manzanas.
Fuente	Contiene atributos de la fuente de datos de los censos nacionales, que es una investigación estadística que recopilar, organizar, analizar y difundir los datos captados. Aquí es donde se determina el periodo de ejecución (Fecha o periodo de referencia).
Vivienda	Contiene todos los atributos propios de una vivienda en base al CPV 2007, está al nivel atómico. Donde vivienda es una edificación construida, adaptada o convertida para ser habitada por una o más personas en forma permanente o temporal.
Hogar	Contiene todos los atributos propios de un hogar en base al CPV

Dimensión	Descripción
	2007, está al nivel atómico. El hogar lo conforma la persona o conjunto de personas, sean parientes o no, que ocupan en su totalidad o en parte una vivienda.
Población	Contiene todos los atributos propios de una persona en base al CPV 2007, está al nivel atómico.
Comunidad Indígena	Contiene atributos de las comunidades indígenas y de los pueblos étnicos.

Tabla 12: Dimensiones del DW de los Censos Nacionales

- Perfil de Dimensiones

En Tabla 14 se muestra los perfiles de dimensiones, son aquellas dimensiones donde existen muchos perfiles que coinciden. En otras palabras estas dimensiones PROFILES son las **agregadas** de las dimensiones para vivienda, hogar y población.

Dimensión	Descripción
PERFIL__POBLACION	Perfil de Dimensión es la tabla agregada de la dimensión población
PERFIL_HOGAR	Perfil de Dimensión es la tabla agregada de la dimensión hogar
PERFIL_VIVIENDA	Perfil de Dimensión es la tabla agregada de la dimensión vivienda

Tabla 13: Perfil de Dimensiones del DW de los Censos Nacionales

- Tabla de Hechos a Nivel Atómico

La solución a nivel atómico (mínimo nivel) cuenta con tres tablas de hechos tales como vivienda, hogar y población, como se muestra en la Tabla 15 siguiente:

Dimensión	Descripción
FACT_POBLACION	Tabla de hechos a nivel atómico que contiene las medidas relacionadas a la población
FACT_HOGAR	Tabla de hechos a nivel atómico que contiene las medidas relacionadas al hogar
FACT_VIVIENDA	Tabla de hechos a nivel atómico que contiene las medidas relacionadas a la vivienda

Tabla 14: Tabla de Hechos a nivel atómico del DW de los Censos Nacionales

- Tabla de Hechos a Nivel Agregadas

En esta sección se hace uso de las agregadas para las tablas de hechos, en donde entra a tallar los PROFILES de dimensiones (PF_VIVIENDA, PF_HOGAR, PF_POBLACION), quienes son los que se relacionaran a estas agregadas.

A continuación en Tabla 16 se muestran las agregadas a nivel de manzana.

Dimensión	Descripción
AGREGADA_VIVIENDA_MZA	Tabla agregada para vivienda a nivel de manzana.
AGREGADA_HOGAR_MZA	Tabla agregada para hogar a nivel de manzana.
AGREGADA_POBLACION_MZA	Tabla agregada para población a nivel de manzana.
AGREGADA_VIVIENDA_DISTRITO	Tabla agregada para vivienda a nivel de distrito.
AGREGADA_HOGAR_DISTRITO	Tabla agregada para hogar a nivel de distrito.

Tabla 15: Tabla de Hechos a nivel agregado del DW de los Censos Nacionales

- Tablas Lookups

Son pequeñas tablas de pocos registros que estarán relacionados a sus dimensiones respectivas como son vivienda, hogar y población. En el modelo

físico se detalla cada una de estas pequeñas tablas. A continuación en Tabla 17 se listan las tablas lookup asociadas a cada dimensión:

Tabla	Descripción
VIVIENDA	Que contiene las descripciones de las características de las viviendas, como tenencia de artefactos eléctricos entre otros.
HOGAR	Que contiene las descripciones de las características de las hogares, como : Tipo de vivienda, Condición de ocupación, Pared, etc.
POBLACION	Que contiene las descripciones de las características de la población, como Parentesco, Sexo, Edad, Educación, etc.

Tabla 16: Tabla de Lookups del DW de los Censos Nacionales

4.4 Procesos ETL

Para trasladar los datos desde las fuentes operacionales al DW es necesario realizar los procesos de extracción, transformación y carga de datos.

El desarrollo de esta actividad requiere especial cuidado pues ha sido necesario la corrección de algunos datos, cuyos valores a nivel macro (nacional, departamental, provincial y distrital) no presentaban problemas, pero cuando se llegaba a niveles menores como centro poblados, zonas y manzanas mostraban resultados inconsistentes. Para esta actividad se utilizo el software SPSS (Statistical Package for the Social Sciences) para identificar los errores y hacer las correcciones, teniendo en cuenta que los valores totales no podrían ser modificados pues estos están publicados como resultados oficiales.

Otro aspecto que consumió gran cantidad de tiempo es que la información no estaba estandarizada y no contaba con información técnica asociada (metadatos).

Para realizar los procesos de ETL se utilizo el Open Source Kettle denominado también como el Pentaho Data Integration conocido como Keetle.

En Figura 32 se muestra el esquema general del proceso ETL realizado.

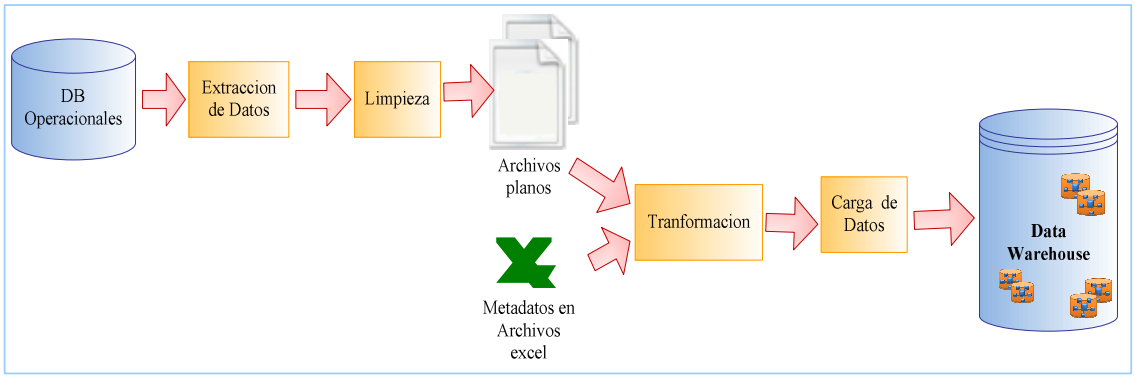

Figura 32: Procesos ETL

4.5 Desarrollo de aplicación para presentación de los dato al usuario

A partir del un modelo multidimensional creado y haciendo uso de la herramienta OLAP (online analytical processing) se ha elaborado el aplicativo denominado "Sistema de Difusión de Los Censos Nacionales" el cual muestra reportes y gráficos flexibles y dinámicos según las necesidades de los:

- Tomadores de decisiones del gobierno central
- Tomadores de decisiones de gobiernos regionales y locales
- Analistas de información de empresas privadas
- Académicos
- Profesionales
- Estudiantes y usuarios diversos

Cada reporte cuenta con metadatos o fichas técnicas que describen las características de los datos, lo cual ayuda al entendimiento de la información de manera rápida y simple.

Así mismo, los reportes generados a partir del DW, mostraran la información a todos los niveles geográficos: nacional, departamental, provincial, distrital, por área urbana y rural y centros poblados.

En la dirección http://inei.gob.pe/ineidw se muestra el aplicativo de acceso a la información del DW de los censos cuya pantalla principal se muestra en la Figura 33.

Figura 33: Pantalla principal del DW del INEI

Donde se observa que ya se tiene integrado tres censos: Censos de Población y Vivienda 2007, Censo Nacional de Comunidades Indígenas y el Censo Nacional Económico 2008.

En la Figura 34, se muestra la los cuadros estadísticos disponibles del tema Población.

Figura 34: Lista de cuadros estadísticos de población

Donde se muestra la lista de cuadros estadísticos disponibles y cada título muestra una breve descripción de su contenido.

También se aprecia que cuenta con un metadato que describe en forma detalla las características del contenido del cuadro, como se muestra en la Figura 35.

Ficha Técnica del Cuadro Estadístico Predefinido	
Información del Cuadro Estadístico Predefinido	
Tema de Cuadro Estadistico	Población
Titulo Corto del Cuadro	POBLACIÓN TOTAL, POR GRANDES GRUPOS DE EDAD, SEGÚN SEXO Y TIPO DE VIVIENDA
Descripción	Contiene información sobre la Población total censada, clasificada por sexo, grandes grupos de edad según área urbano-rural, departamento, provincia y distrito y tipo de vivienda.
Especificaciones Técnicas	Población total: En números absolutos, por grandes grupos de edad: menores de 1 año, de 1 a 14, 15 a 29, 30 a 44, 45 a 64, 65 a más años. Vivienda Particular: Casa independiente, Departamento en edificio, Vivienda en quinta, Vivienda en casa de vecindad (callejón, solar o corralón), Choza o cabaña, Vivienda improvisada, Local no destinado para habitación humana, Otro tipo. Vivienda Colectiva: Hotel, hostal, hospedaje, Casa pensión, Hospital, clínica, Cárcel, centro de readaptación social, Asilo, Aldea infantil, orfelinato, etc, Otro. Otro tipo: En la calle, personas sin vivienda, garita, puerto, aeropuerto, etc.
Nivel Geográfico	La información de este cuadro se puede desagregar a nivel nacional, departamental, provincial, distrital y de centro poblado.
Fuente	INEI - Censos Nacionales 2007: XI de Población y VI de Vivienda
Nota	Este sistema proporciona la información de los censos nacionales a nivel nacional, departamental, provincial, distrital y centro poblados, facilitando realizar análisis dinámicos con múltiples dimensiones (tiempo, espacio y diversas temáticas) para la toma de decisiones. No se empadronó a la población del distrito de Carmen Alto, provincia de Huamanga, departamento de Ayacucho
Variable: Tipo de vivienda - Grupo	
Información de Variables	
Grupo de Variable	Características de la Vivienda
Titulo Corto de la variable	Tipo de vivienda - Grupo
Título Largo de Variable	Tipo de vivienda - Grupo
Definición de la variable	Vivienda es una edificación o unidad de edificación independiente construida, adaptada o convertida para ser habitada por una o más personas en forma permanente o temporal. Los tipos de viviendas son: viviendas particulares, viviendas colectivas y viviendas de otro tipo.

Figura 35: Metadatos de la fuente de información

Para mostrar las funcionalidades dinámicas de los cuadros estadísticos se seleccionara el "Cuadro 2: POBLACIÓN TOTAL, POR GRANDES GRUPOS DE EDAD, SEGÚN SEXO Y TIPO DE VIVIENDA" que se muestra en la Figura 36.

POBLACIÓN TOTAL, POR GRANDES GRUPOS DE EDAD, SEGÚN SEXO Y TIPO DE VIVIENDA

FILTROS BÁSICOS: Departamento: Total Provincia: Total Distrito: Total

Área: Total

Nº Filas: 10 | Nº

Tipo de Vivienda	Sexo	Medidas Grandes Grupos de Edad	Numero de Personas					
			Total	Menos de 1 Año	1 a 14 Años	15 a 29 Años	30 a 44 Años	45 a 64 Años
Total			27,412,157	500,672	7,856,861	7,554,204	5,588,602	4,147,131
Vivienda Particular	Total		27,057,199	496,798	7,831,933	7,408,812	5,480,633	4,091,262
	Hombre		13,361,977	252,422	3,992,000	3,646,933	2,644,069	1,990,727
	Mujer		13,695,222	244,376	3,839,933	3,761,879	2,836,564	2,100,535
Vivienda Colectiva	Total		295,792	3,571	20,546	123,164	88,855	44,797
	Hombre		219,350	1,970	10,738	94,910	71,240	32,788
	Mujer		76,442	1,601	9,808	28,254	17,615	12,009
Otro tipo	Total		59,166	303	4,382	22,228	19,114	11,072
	Hombre		41,313	145	2,319	14,909	14,259	8,268
	Mujer		17,853	158	2,063	7,319	4,855	2,804

Nota: Este sistema proporciona la información de los censos nacionales a nivel nacional, departamental, provincial, distrital y ce facilitando realizar análisis dinámicos con múltiples dimensiones (tiempo, espacio y diversas temáticas) para la toma de de empadronó a la población del distrito de Carmen Alto, provincia de Huamanga, departamento de Ayacucho.

Figura 36: Cuadros estadísticos dinámicos sobre población

En grafico anterior, el cuadro estadístico cuenta con opciones "Filtros Básicos", que permite realizar selecciones según los atributos disponibles. Cualquier objeto del cuadro puede pasar a esta zona (arrastrar con el mouse el objeto del cuadro). Igualmente cualquier objeto que se encuentre en esta zona puede pasar a filas o columnas del cuadro. Esta acción lo realizara el usuario a la medida de sus necesidades.

Otras funcionalidades importantes a mencionar son:

- Las flechas ubicadas en las cabeceras de las columnas permiten ordenar de manera ascendente o descendente la información
- Permite permutar las filas por columnas y columnas por filas.
- Permite cambiar el orden las columnas o filas
- Dispone del modo identado, que es un efecto para mostrar un cuadro en forma compacta sin niveles o quiebres.
- Permite eliminar atributos del cuadro estadístico.
- Dispone de funciones estadísticas para ser insertar según necesidad del usuario.

En Figura 37 se muestra cuadro estadístico personalizado, donde “Grandes Grupos de Edad” ha pasado a la zona de “Filtros Básicos”, “Sexo” ha sido reubicado como columna y el modo identado que está disponible para “Tipo de Vivienda” que ha sido desplegado y muestra las categorías correspondientes.

POBLACIÓN TOTAL, POR GRANDES GRUPOS DE EDAD, SEGÚN SEXO Y TIPO DE VIVIENDA

FILTROS BÁSICOS: Departamento: Total Provincia: Total Distrito: Total
Área: Total Grandes Grupos de Edad: Total

Nº Filas: 10 | Nº Columnas: 3

Tipo de Vivienda	Tipo de Vivienda	Sexo / Medidas	Total Numero de Personas	Hombre Numero de Personas	Mujer Numero de Personas
Total			27,057,199	13,361,977	13,695,222
Vivienda Particular	Total		27,057,199	13,361,977	13,695,222
	Casa independiente		23,880,623	11,796,264	12,084,359
	Departamento en edificio		1,144,856	535,221	609,635
	Vivienda en quinta		435,187	211,965	223,222
	Vivienda en casa de vecindad		354,431	174,343	180,088
	Choza o cabaña		886,504	464,398	422,106
	Vivienda improvisada		296,861	147,762	149,099
	Local no destinado para habitación humana		30,931	17,772	13,159
	Otro tipo particular		27,806	14,252	13,554

Nota: Este sistema proporciona la información de los censos nacionales a nivel nacional, departamental, provincial, distrital y centro poblados, facilitando realizar análisis dinámicos con múltiples dimensiones (tiempo, espacio y diversas temáticas) para la toma de decisiones. No se empadronó a la población del distrito de Carmen Alto, provincia de Huamanga, departamento de Ayacucho.

Figura 37: Cuadro estadístico dinámico de población y vivienda

- También permite generar gráficos estadísticos totalmente dinámicos, donde funcionan los filtros y todo cambio que se efectué con el cuadro estadístico. En ejemplo de Figura 38 se muestra grafico estadístico de la población en viviendas particulares por grupos de edad en distrito de Punta Hermosa de la provincia de Lima del departamento de Lima.

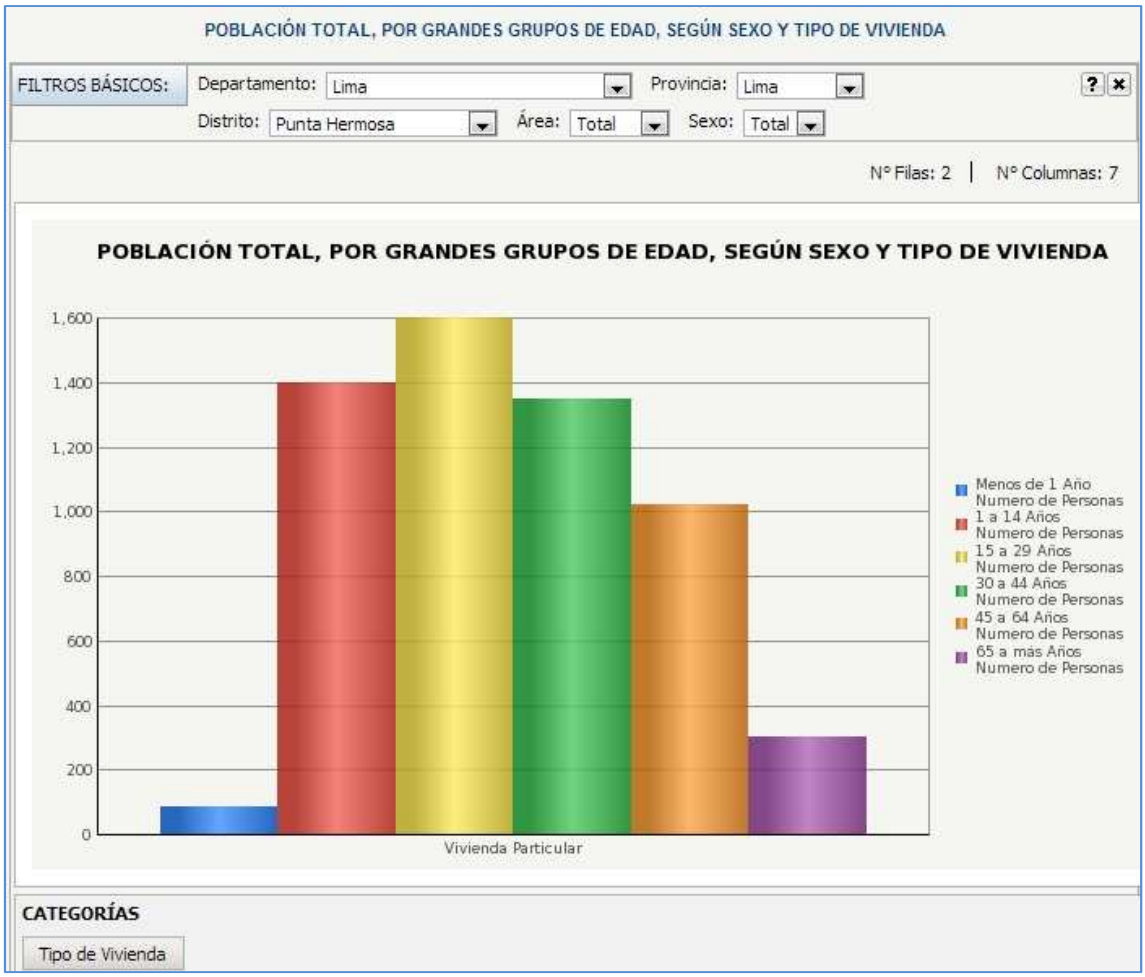

Figura 38: Gráficos dinámicos de población y vivienda

- Para el mismo distrito seleccionado, como elemento de análisis se ha movido a filas la categoría "Sexo", luego se ha cambiado el tipo de grafico de "Barra Vertical" a "Barra horizontal", lo cual permite realizar un análisis comparativo de la población por rangos de edad que residen en las viviendas particulares del distrito de Punta Hermosa, como se muestra en Figura 39.

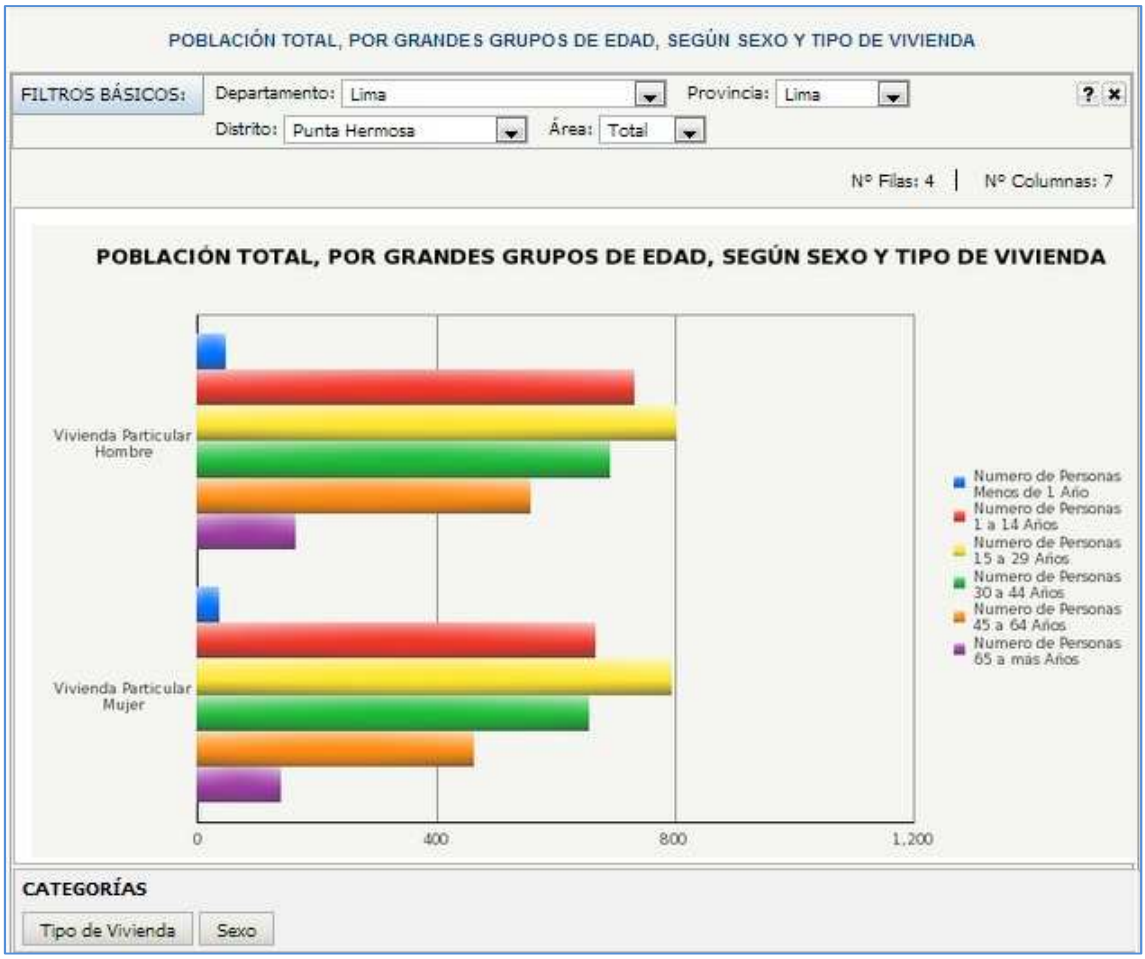

Figura 39: Gráfico dinámico de población según categoría de vivienda

- Para el mismo distrito seleccionado, como elemento de análisis se ha quitado de fila la categoría "Sexo" el que se ha movido a la zona de "Filtros Básicos", luego se ha cambiado el tipo de grafico de "Barra horizontal" a "Circular", lo cual permite realizar un análisis de la población total por rangos de edad que residen en las viviendas particulares del distrito de Punta Hermosa, como se muestra en Figura 40.

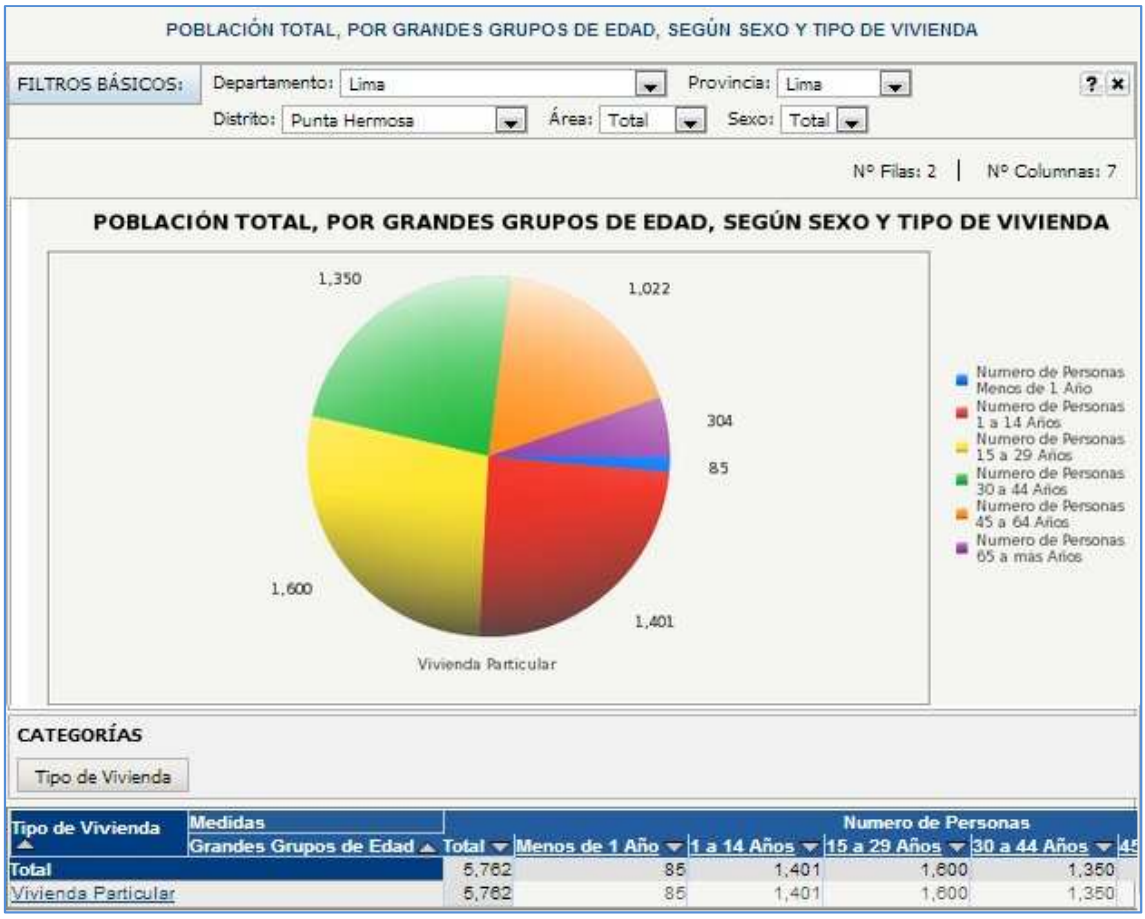

Figura 40: Gráfico dinámico de población según tipo de vivienda

- Otra funcionalidad muy valorada por los usuarios es el análisis espacial que pueden realizar haciendo uso de mapas.

  Por ejemplo en Figura 41 se muestra un mapa temático del indicador "Tasa de escolarización de la población de 6 a 11 años" a nivel departamental. Para lo cual se selecciono la fuente "Censos Nacionales 2007", el Tema "Social", Subtema "Educación" y en Indicador "Tasa de escolarización de la población de 6 a 11 años". Al ejecutar la opción mapa "Departamental", se mostrará el mapa temático siguiente:

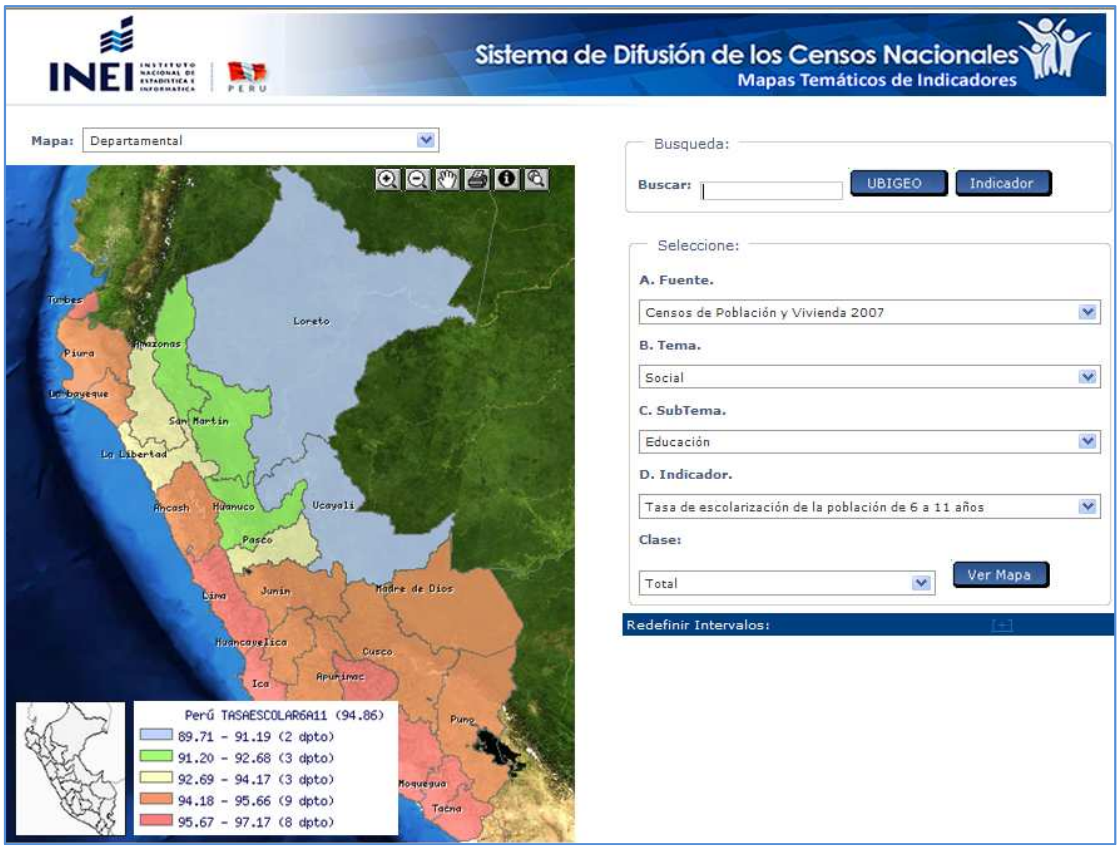

Figura 41. Análisis espacial de indicadores

Donde la leyenda del mapa temático se muestra el valor del indicador a nivel nacional y por defecto cinco intervalos, en cada uno de los cuales se indica el número de casos. Los valores de los intervalos pueden ser modificados según las necesidades del usuario.

4.6 Implementación

Para esta actividad se verifico la existencia de las condiciones necesarias; en lo referido a los datos, al hardware, software y comunicaciones que permita el correcto funcionamiento del Data Warehouse y el aplicativo de consulta para el usuario final.

Así mismo, se realizaron capacitaciones y mentarías para la transferencia de conocimiento al equipo permanente de la institución, quienes brindaran el soporte técnico y puedan dar solución ante cualquier eventualidad desfavorable al correcto funcionamiento del Data Warehouse de los Censos.

4.7 Mantenimiento y Crecimiento

El DW de los censos nacionales fue construido e implementado considerando que sea el soporte para la toma de decisiones del Sistema Estadístico Peruano, por lo cual el mantenimiento y crecimiento del mismo es un proceso continuo que se ha incorporado a las actividades del INEI.

Para tal fin se conformo un equipo de trabajo que se encargue de su monitoreo permanente y se encargue de darle el mantenimiento o soporte técnico necesario para asegurar su crecimiento continuo.

Es así que ya se ha logrado incorporar nuevas fuentes de datos al DW de los Censos Nacionales, estas son el Censo Nacional Económico y Registro Nacional de Municipalidades, el Censo Nacional 1993 y se viene trabajando el Censo Nacional Agropecuario.

4.8 Diseño de la Arquitectura Técnica

La plataforma tecnológica del DW de los Censos proporciona una herramienta de gestión que permite acceder a grandes volúmenes de datos, los cuales pueden ser desplegados en tiempo real. Se basa en un sistema multicapas distribuido sobre cinco servidores tipo Blade, que aseguran el crecimiento futuro del DW, como se muestra en Figura 42.

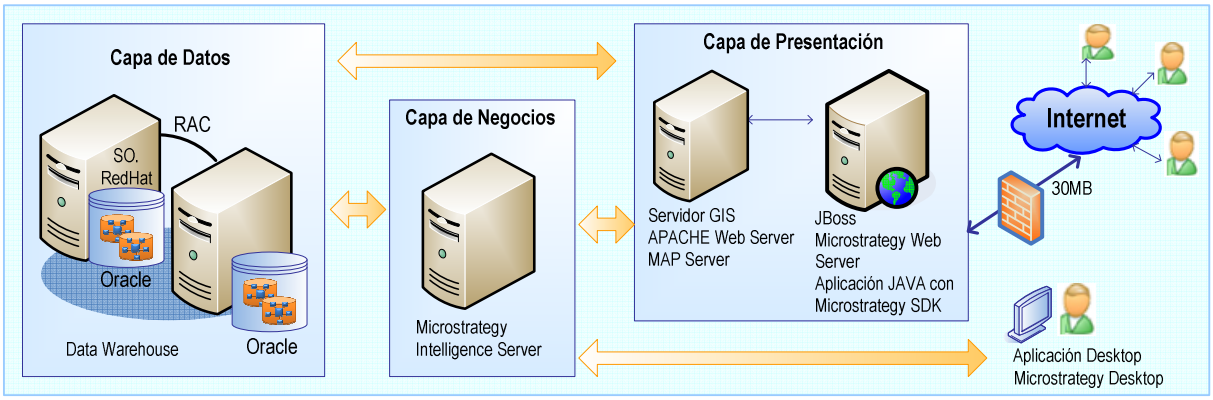

Figura 42: Plataforma tecnológica del DW

Donde:

- Capa de Datos, cuenta con dos servidores configurados en espejo utilizando Real Application Cluster (RAC) del manejador de Base de Datos Oracle, lo que asegura la alta disponibilidad del servicio, el balanceo de carga y su escalabilidad, lo que ofrece un sistema que continuará trabajando ante la caída de alguno de sus nodos.

- Capa de Negocio, cuenta con un servidor donde se encuentra configurado el motor de OLAP: MicroStrategy Intelligence Server que se encarga de gestionar las consultas realizadas por los usuarios siguiendo las reglas de negocio.

- Capa de presentación, cuenta con un servidor de aplicaciones JBoss sobre el que se encuentra configurado el servidor Web MicroStrategy WebServer, con el aplicativo desarrollado en Java e integrado con componentes de MicroStrategy SDK. También se cuenta con un servidor GIS con Apache y el software MapServer.

Capitulo 5: CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

1) La presente tesina presenta el desarrollo de un Data Warehouse de los Censos Nacionales para el Instituto Nacional de Estadísticas e Información (INEI) de Perú, que es el organismo que produce y difunde la estadística oficial del país. Dicho DW integra la información clave para usuarios internos (directivos e investigadores) y externos (funcionarios de entidades públicas y privadas) del INEI quienes requieren conocer y evaluar la situación económica, social y demográfica del país para tomar decisiones en pos de su bienestar.

2) Para realizar en forma exitosa un proyecto de desarrollo de DW ha sido vital que los altos funcionarios tomadores de decisión del INEI participen en forma activa en el proyecto, viéndolo no como un tema netamente técnico si no como una solución estratégica para el posicionamiento del instituto como una entidad transparente que difunde la información censal con calidad a todo nivel geográfico (departamento, provincia, distrito, centros poblados) e incluso nivel atómico (zonas y manzanas censales).

3) Conocer las necesidades de información de todos los usuarios, ha sido determinante para la construcción del modelo multidimensional y la elaboración de los cuadros estadísticos para la presentación de los datos, con toda la flexibilidad requerida de tal modo que el usuario pueda construir los cuadros estadísticos a medida de sus necesidades. Así mismo, el DW abre una ventana de posibilidades que impulsa el trabajo colaborativo, pues brinda

nuevas herramientas de análisis al personal de las diferentes áreas que trabajan en el INEI.

4) El estudio comparativo de las metodologías de Ralph Kimball, Bill Inom y Ramón Barquín, ha permitido optar por la que brinda mayor flexibilidad y simplicidad para la construcción de DW con información censal, la cual a diferencia de los DW comerciales no tiene una frecuencia de actualización continua, ya que los periodos censales son cada cinco (5) o diez (10) años. Otro factor determinante para optar por esta metodología fue que se contó con la asistencia técnica directa de Ramón Barquín.

5) Un aspecto critico y que tomó tiempo es la estandarización de la información resultante de los sistemas operacionales, para lo cual se conto con equipo de trabajo de especialistas temáticos, quienes se encargaron de elaborar los metadatos de la información censal, de los cuadros estadísticos, de los indicadores, entre otros. Esto fue un gran aporte al proceso productivo de información censal, ya que los lineamientos y estándares aplicados a este proyecto han sido aplicados a todos los proyecto posteriores, de tal modo que así se asegura la integración de la información.

6) Los procesos de extracción, transformación y carga permitieron detectar inconsistencias a nivel atómico lo que no era visible a nivel total, ello obligó a realizar procesos de limpieza de datos para corregirlos aplicando el procedimiento para el control de calidad así como herramientas de TI a fin de alcanzar la exactitud para toda la data del censo. Estos procedimientos quedan como lecciones aprendidas para controlar la calidad de los datos de los futuros censos que se carguen al DW.

7) Se optimizo el tiempo de atención a los requerimientos de cuadros estadísticos con información de los censos nacionales, el usuario ya no tiene que acercarse a las oficinas del INEI a dejar solicitudes de información y esperar pasivamente la respuesta. Ahora el usuario accede a la información censal mediante la aplicación en Internet y puede el mismo hacer sus cuadros exactamente como los necesita. En consecuencia los costos de acceder a la información censal ha sido eliminado.

8) Presentación de cuadros estadísticos totalmente dinámicos y personalizables, con lo cual el set de 87 cuadros estáticos presentados como resultado definitivo de cada censo de población y vivienda, se ha convertido en millones de combinaciones de cuadros estadísticos.

5.2 Recomendaciones

1) Para el desarrollo exitoso de un proyecto de esta índole es vital el liderazgo de la máxima autoridad, quien debe tener claro que el DW es un recurso estratégico y táctico para la institución y debe promover la participación de los usuarios que aporten al desarrollo. Así mismo, es necesario una adecuada asignación de recursos humanos y tecnológicos, para lo cual la alta dirección debe gestionar la obtención de los mismos.

2) Para desarrollo de proyecto de DW es necesario la conformación de equipos de trabajo multidisciplinarios (especialistas en TI y especialistas temáticos), así mismo es vital contar con un comité institucional quienes estén validando que se estén cumpliendo los objetivos del proyecto.

3) Para el desarrollo de DW es necesario la aplicación de una metodología de trabajo acorde a las características y necesidades propias de la institución.

4) Considerando que el INEI produce censos y encuestas, se recomienda la construcción de un Datamart para las encuestas periódicas que se realizan y que actualmente están en bases de datos independientes y con aplicaciones de presentación de información en diversos formatos.

5) El INEI como ente rector del Sistema Estadístico Nacional debe impulsar el crecimiento continuo del DW de los Censos Nacionales para ser el gran centro de información estadística de todas las entidades del sector público, en el que se integre la información estadística de los sectores.

REFERENCIAS BIBLIOGRAFICAS

Libros

[Barquin 1997] Ramon Barquin y Herbert Edelstein. "The Building, Using, and Managing The Data Warehouse". Prentice Hall PTR., 2004.

[Basandra 2013] Suresh Basandra. "Database, Data Warehouse and Business Intelligence. JOB Interview , Questions and Answers". Basandra Publications 2013).

[Bernabeu 2010] Bernabeu Ricardo Dario. "Data Warehousing: Investigación y Sistematización de Conceptos. HEFESTO: Metodología para Construcción de un Data Warehouse". Córdoba, Argentina 2010.

[Cano 2007] Josep Lluis Cano. "Inteligencia de Negocios: Competir con Información". ESADE, 2007.

[CIBERTEC 2011] CIBERTEC. "Inteligencia de Negocios, Teoria". CIBERTEC, 2011.

[Harjinder 1996] Harjinder S. Gill y Prakash C. Rao. "Data Warehousing: La Integración de la Información para la Mejor Toma de Decisiones".Prentice Hall Hispanoamérica, 1996.

[Kimball 2004] Raplh Kimball y Joe Caserta. "The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Clearing, Conforming, and Delivering Data". Willey Publishing, Inc., 2004.

[Medina 209] Edison Medina La Plata. "Bisiness Intelligence". UPC, 2009.

[Parra 1998] Enrique Parra Iglesias. "Tecnologías de la Información en el Control de Gestión". Diaz de Santos 1998.

[Peña 2006] Alejandro Peña Ayala. "Inteligencia de Negocios: Una Propuesta para su Desarrollo en las Organizaciones". Instituto Politécnico Nacional, 2004.

[Tompkins 1998] James A. Tompkins, Jerry D. Smith. "The Data Warehouse Managemente Handbook".

[Torben 2010] Torben Bach Pedersen, Mukesh K. Mohania y A Min Tjoa. "Data Warehousing and Knowledge Discovery. 12th International Conference, DaWak 2010". Springer-Verlag Berlin Heidelberg 2010.

Direcciones electrónicas

Ampliación de conceptos del Modelado Dimensional. Roberto Espinosa, 2010
http://churriwifi.wordpress.com/2010/04/19/15-2-ampliacion-conceptos-del-modelado-dimensional/

Consulta Interactiva de Datos en la Pagina Web del INEGI en Internet. Antecedentes e Importancia.
http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=1552

Data Warehouse para la Prestación del Servicio Público de Información Estadística. Ricardo Lujan Salazar. Instituto Nacional de Estadística, Geografía e Informática (INEGI). México, 2003.
http://www.eclac.cl/deype/mecovi/docs/TALLER11/15.pdf

Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon models.
Mary Breslin, 2004
http://olap.it/Articoli/Battle%20of%20the%20giants%20-%20comparing%20Kimball%20and%20Inmon.pdf

Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models.
Mary Breslin, Mayo 2007
http://www.tdan.com/view-articles/4768

Data Warehouse
Cristhian Herrera. Octubre 2007.
http://www.adictosaltrabajo.com/tutoriales/tutoriales.php?pagina=datawarehouse

El INE y su producción estadística: una nota histórica sobre los últimos 50 años.
Mariano Gómez del Moral, Noviembre 2011
http://www.ine.es/ss/Satellite?L=es_ES&c=INEDocTrabajo_C&cid=1259937252474&p=1254735839320&pagename=MetodologiaYEstandares%2FINELayout

FIELD LISTING : INTERNET USERS. Agencia Central de Inteligencia
https://www.cia.gov/library/publications/the-world-factbook/fields/2153.html

Glosario de Gartner, enero 2006.
www.gartner.com

La aportación del 'Censo de Viviendas 2001' al análisis de la problemática de la vivienda en España.
Eduardo Teijeiro Alfonsín. Subdirector general adjunto de Censos y Padrón. Mayo 2004.
http://dialnet.unirioja.es/servlet/articulo?codigo=951241

ONGEI: Celebro el Día de Internet, 2012. Oficina Nacional de Gobierno Electrónico del Perú.
http://www.ongei.gob.pe/noticias/ongei_noticias_detalle.asp?pk_id_entidad=1878&pk_id_noticia=370

Técnicas de difusión de grandes bases de datos. La difusión censal abre una nueva ventana en las bases de datos del sistema estadístico público.
Armando de la Torre, Jefe de Área Editorial Instituto Nacional de Estadística. Setiembre 2004
http://www.ief.es/documentos/investigacion/seminarios/estadistica_publica/2004_16Septiembre.pdf

Presupuesto del Censo Nacional 2007. Pagina web del INEI.
http://www1.inei.gob.pe/web/Transp_Admin/attach/6848.pdf

Primeros Resultados del Censo Nacional de Población y Vivienda 2007
http://censos.inei.gob.pe/censos2007/documentos/Resultado_CPV2007.pdf

Sociedad de Estadística e Investigación Operativa, Boletín de la SEIO, Volumen 22, número 3, JULIO 2006
http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/PUBLICACIONES/BoletinVol22Num3.pdf

Tesis

Córdova Rodríguez, Diana Julissa. "Desarrollo de un Data Mart para Incrementar la Capacitación de Clientes en una Empresa de Telecomunicaciones". Tesina para optar el título profesional de Ingeniero de Sistemas. Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos, Lima 2011.

Rojas, Mariana Isabel. "Monografía de Adscripción: Data Warehouse". Para optar el titulo de Licenciatura en Sistemas de Información. Facultad de Ciencias Exactas y Naturales y Agrimensura. Universidad Nacional del Nordeste. Argentina 2009.

Ibarra, María de los Ángeles. "Monografía de Adscripción: Procesamiento Analítico en Línea (OLAP)". Para optar el titulo de Licenciatura en Sistemas de Información. Facultad de Ciencias Exactas, Naturales y Agrimensura. Universidad Nacional del Nordeste. Argentina 2006